Zuverlässigkeit von CMOS-Bauelementen auf SOI für den Betrieb bei 250 °C

Von der Fakultät für Ingenieurwissenschaften
Abteilung Elektrotechnik und Informationstechnik
der Universität Duisburg-Essen

zur Erlangung des akademischen Grades

Doktorin der Ingenieurwissenschaften

genehmigte Dissertation

von

Katharina Grella
aus
Wiesbaden

1. Gutachter:	Prof. Dr.-Ing. Holger Vogt
2. Gutachter:	Prof. Dr.-Ing. Ulrich Hilleringmann
Tag der mündlichen Prüfung:	09.07.2013

Inhaltsverzeichnis

Danksagung

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftliche Mitarbeiterin am Fraunhofer-Institut für Mikroelektronische Schaltungen und Systeme (IMS). Zu ihrem Gelingen haben viele Personen beigetragen, denen ich hier danken möchte.

Zunächst gilt mein Dank Prof. Dr.-Ing. Holger Vogt für die Themenstellung, die Betreuung dieser Arbeit und das Feedback während der Doktorandenvorträge. Prof. Dr.-Ing. Ulrich Hilleringmann danke ich für die Übernahme des Zweitgutachtens.

Bei Dr. Uwe Paschen möchte ich mich für die Unterstützung während der gesamten drei Jahre bedanken. Unsere freitäglichen Diskussionen über die Messungen und ihre mögliche Interpretation waren mir eine große Hilfe.

Danken möchte ich auch Dr. Stefan Dreiner für die vielfältigen und hilfreichen Diskussionen und sein großes Interesse an den unterschiedlichen Themen dieser Arbeit.

Weiterhin möchte ich Dirk Dittrich für die Unterstützung bei messtechnischen und H10-spezifischen Fragen und Dr. Miriam Klusmann für die erfolgreiche Betreuung der H10-Wafer danken. Mein Dank gilt auch Holger Kappert und Alexander Schmidt für die angenehme Zusammenarbeit im Hochtemperatur-Projekt und die Diskussionen über Ringoszillatoren und Bandgap-Schaltungen. Wolfgang Heiermann möchte ich für den Aufbau der Chips danken. Und natürlich darf auch mein Bürokollege Janusz Pieczynski nicht vergessen werden, dem ich für die Hilfestellung bei messtechnischen Problemen, die Diskussionen über Zuverlässigkeit und die gute Büroatmosphäre dankbar bin.

Ich danke auch all meinen anderen Kollegen der Abteilung CTB für nette Unterhaltungen in den Mittagspausen, Kuchen, Bowling-Abende und vieles mehr. Die angenehme Arbeitsatmosphäre hat dazu beigetragen, dass ich mich in der Abteilung sehr wohl gefühlt habe.

Ein Extra-Dank an alle, die beim Korrekturlesen mitgewirkt haben.

Dank gebührt auch meiner Familie, die mich während des Studiums und der Promotion immer unterstützt hat. Und natürlich danke ich auch Staffan, für sein offenes Ohr und die Kraft, die er mir gegeben hat, so dass ich auch bei Rückschlägen den Mut nicht verloren habe.

Zusammenfassung

Diese Arbeit behandelt die Zuverlässigkeit eines 1 µm-CMOS-Prozesses, der auf Silicon-on-Insulator (SOI)-Wafern gefertigt wurde und für die Anwendung im Temperaturbereich von -40 °C bis 250 °C vorgesehen ist. Dabei werden zwei Themen miteinander verbunden, die Hochtemperaturelektronik und die Zuverlässigkeit. Von Hochtemperaturelektronik spricht man, wenn es um die Anwendung von integrierten Schaltungen im Temperaturbereich oberhalb von 125 °C geht. Da die meisten Zuverlässigkeitsmechanismen, die Halbleiter betreffen können, durch Temperatureinfluss beschleunigt werden, ist die Zuverlässigkeit eines solchen Hochtemperaturprozesses besonders kritisch. Zudem ist die Mehrzahl der häufig angewandten Methoden zur beschleunigten Messung und Bewertung verschiedener Zuverlässigkeitsaspekte nicht für Testtemperaturen oberhalb von 250 °C vorgesehen. In dieser Arbeit wurde deshalb evaluiert, welche Aspekte bei Temperaturen von 250 °C und darüber hinaus besonders relevant sind und ob zusätzliche Mechanismen die Zuverlässigkeit beeinflussen.

Die vorgestellten Untersuchungen beziehen sich auf das so genannte End-of-Line Monitoring, d. h. auf die Beurteilung der Zuverlässigkeit einzelner Bauteile nach der Prozessierung. Im Detail sind dies die Bewertung der Qualität und Zuverlässigkeit des 40 nm dicken Gateoxids und der Wolframmetallisierung, die Untersuchung der Programmierbarkeit, des Datenerhalts und der Zyklenfestigkeit von Single-Poly-EEPROM-Speicherzellen sowie die Charakterisierung und die Untersuchung der Langzeitstabilität von Transistoren und Testschaltungen.

Bei EEPROM-Speicherzellen tritt oberhalb von 250 °C ein Mechanismus in Erscheinung, der für einen zusätzlichen Datenverlust und damit eine erhöhte Aktivierungsenergie sorgt. Bei der Zuverlässigkeit der Metallisierung spielt Elektromigration keine Rolle, bei Stressmigrationstests an Polysilizium-Metall-Kontakten konnte aber ein mit der Zeit ansteigender Widerstand beobachtet werden, der eine Folge aufgebrochener Wasserstoffbindungen ist. Bei Messungen zur Zuverlässigkeit des Gateoxids wurde festgestellt, dass Elektroneninjektion von Seiten des Polysilizium-Gates einen negativen Einfluss auf Durchbruchspannungen und Durchbruchzeiten von Gateoxidkondensatoren hat. Bei Untersuchungen zu den Auswirkungen heißer Ladungsträger auf die Lebensdauer von NMOS- und PMOS-Transistoren wurde der Einfluss des Back-Gates deutlich. Bei Messungen zur Parameterinstabilität von Transistoren bei negativer Gate-Source-Spannung (NBTI) konnte eine alternative Methode erfolgreich angewandt werden, bei der man statt der Schwellenspannungsänderung eine Stromänderung misst und so das Problem der sonst auftretenden Ausheilung umgeht. Bei der Charakterisierung von Transistoren und Ringoszillatoren wurde eine prinzipielle Funktionalität bis 400 °C festgestellt. Zudem traten im Rahmen der Messgenauigkeiten in einem Zeitraum von einigen Tausend Stunden keine Parameteränderungen an Ringoszillatoren und Bandgap-Schaltungen auf.

Die Verwendung von CMOS auf SOI ist prinzipiell bis 400 °C möglich. Durch diese Arbeit ist eine Grundlage gelegt, um beschleunigte Zuverlässigkeitsuntersuchungen auch bei Temperaturen oberhalb von 250 °C durchführen und damit in einem angemessenen Zeitraum sinnvolle Aussagen über die Zuverlässigkeit bei Einsatztemperaturen von 250 °C machen zu können.

Abstract

This thesis is concerned with the reliability of a 1 µm-CMOS-process, which is fabricated on Silicon-on-Insulator (SOI) wafers and destined to be used at temperatures between -40 °C and 250 °C. It brings together two subjects: high temperature electronics and reliability aspects. In this context, high temperature electronics means the application of integrated circuits at temperatures above 125 °C. As most of the reliability mechanisms which concern semiconductors are accelerated by increases in temperature, the reliability of such a process is especially crucial. Furthermore, the majority of the commonly applied methods for accelerated testing and analysis of the different reliability aspects are not intended to be used at temperatures above 250 °C. In this work, we evaluated which aspects are especially relevant at temperatures of 250 °C and above, and whether additional mechanisms perturb the reliability.

The presented investigations deal with End-of-Line-Monitoring, which means reliability estimation of single devices after their fabrication. Specifically, we consider the assessment of the reliability of the 40 nm gate-oxide and the tungsten-metallization; the analysis of the programmability, the data retention and the endurance behavior of Single-Poly-EEPROM-cells; and the characterization and the long-term-study of transistors and test-circuits.

For EEPROM-cells, a mechanism appears at temperatures above 250 °C which is responsible for additional data loss and high activation energy. Concerning the metallization, electromigration does not play a role, but stress migration tests at polysilicon-metall-contacts showed an increasing resistance with time, which is due to broken hydrogen bonds. Measurements related to gate-oxide reliability indicated that electron injection from the polysilicon-gate has a negative influence on the breakdown-voltage and the time to breakdown of gate-oxide-capacitors. In the investigations of the impact of hot carriers on the lifetime of NMOS- and PMOS-transistors, the back-gate influence became clearly visible. Measurements of the Negative Bias Temperature Instability (NBTI) were successfully carried out by using an alternative method in which current change is measured instead of the variation of the threshold voltage, so that the problem of recovery that would arise otherwise is avoided. Transistors and ring oscillators showed a principal functionality up to 400 °C. In addition to that, no parameter change was detected within some thousands of hours when observing ring oscillators and bandgap references.

The use of CMOS on SOI is in principle possible up to 400 °C. This works establishes a basis for accelerated reliability testing at temperatures above 250 °C in order to obtain information about the reliability at a use temperature of 250 °C in reasonable time.

Kapitel 1

Einleitung

Während der Haupttrend der Mikroelektronik dem Gesetz von Moore hin zu immer kleineren Strukturgrößen und damit größeren Integrationsdichten folgt [MOO65], gibt es noch andere Entwicklungen, die weniger die Integrationsdichte als die Anpassung an spezifische Anforderungen im Blick haben („More than Moore“). Zur zweiten Gruppe zählt der zunehmende Bedarf von Elektronik für raue Umgebungsbedingungen, unter anderem für die Anwendung bei hohen Temperaturen. Von Hochtemperaturelektronik spricht man, wenn die normale Betriebstemperatur einer Schaltung 125 °C oder mehr beträgt [KIR12].

Bedarf für solche Elektronik ist in verschiedenen Branchen vorhanden. Dazu zählen vor allem die Bohrindustrie zur Förderung von Öl oder Gas [PAR03], der Bereich der geothermischen Energiegewinnung [NOR11], der Automobilsektor [JOH04] und die Luft- und Raumfahrt [JOH00], aber auch industrielle Anwendungen im Bereich der Prozessüberwachung und Prozesssteuerung. Tabelle 1.1, die inhaltlich aus [COL97] entnommen wurde, stellt diese Branchen und ihre Temperaturanforderungen gegenüber.

Anwendung	**Temperaturen**
Tiefbohrungen	**75 - 600 °C**
Ölbohrungen	75 - 175 °C
Gasbohrungen	150 - 225 °C
Dampfeinleitung	200 - 300 °C
Geothermische Energie	200 - 600 °C
Automobilsektor	**150 - 600 °C**
Motorraum	-50 - 200 °C
Sensoren am/im Motor	bis 600 °C
Sensoren im Abgassystem	bis 600 °C
ABS (Antiblockiersystem)	bis 300 °C
Luft- und Raumfahrt	**150 - 600 °C**
Interne Ausstattung	150 - 250 °C
Triebwerküberwachung	300 - 600 °C
Oberflächenkontrollen	300 - 600 °C
Satelliten (Venussonde)	150 - 600 °C
Kernkraftwerke	**30 - 550 °C**

Tabelle 1.1: Temperaturanforderungen an Elektronik in einigen Anwendungsbereichen; aus [COL97], S. 247

In den meisten dieser Anwendungsbereiche wird mangels ausreichender Angebote an Hochtemperaturelektronik häufig mit alternativen Konzepten gearbeitet, die aber einige Nachteile haben. So wird beispielsweise herkömmliche Elektronik mit einer aktiven Kühlung ausgestattet, die die Temperaturen moderat hält. Dies erhöht aber den Platzbedarf und den Energieverbrauch und damit die Kosten. Ein anderes Beispiel ist die Trennung von Sensor und Elektronik. Während der Sensor in der wärmeren Umgebung angebracht ist, befindet sich die Elektronik getrennt davon in einem Bereich, in dem niedrigere Temperaturen herrschen. Die Nachteile davon sind große Übertragungswege und damit eine erhöhte Anfälligkeit für Störungen, eine insgesamt wenig kompakte Anordnung und die Notwendigkeit separater Kalibrationen. Elektronik, die bei hohen Umgebungstemperaturen zuverlässig funktioniert, ist daher von großem Vorteil.

Um Elektronik für hohe Umgebungstemperaturen nutzbar zu machen, spielt das Basismaterial der Wafer eine wichtige Rolle. Bei siliziumbasierter Elektronik ist die herkömmliche Bulk-Technologie durch mit der Temperatur stark zunehmende Leckströme der pn-Übergänge nur bis etwa 175 °C einsetzbar. Silicon-on-Insulator (SOI)-Wafer hingegen sind für Hochtemperaturanwendungen gut geeignet, da sich Leckströme hierbei um Größenordnungen reduzieren lassen. Die Möglichkeiten, SOI für Hochtemperaturprozesse zu nutzen, wurden in der Literatur vor allem in den 90er Jahren diskutiert [FRA92], [FLA95], [GEN97] und sind heute wieder aktuell [VAN08]. Alternative Technologien wie Siliziumcarbid-basierte Elektronik sind wegen material- und prozesstechnischer Schwierigkeiten momentan noch keine Lösung.

Neben dem Substratmaterial gibt es noch weitere Aspekte, die bei der Herstellung von hochtemperaturtauglichen Technologien zu beachten sind. Dazu zählt beispielsweise die Metallisierung. Reines Aluminium zeigt bei erhöhter Temperatur schnell elektromigrationsbedingte Ausfallerschienungen, weshalb andere Metalle wie beispielsweise Wolfram für Hochtemperaturanwendungen besser geeignet sind [CHE95]. Außerdem ist die Leistungsfähigkeit von Schaltungen bei hohen Temperaturen stark eingeschränkt. Deshalb sind auch entsprechende Schaltungskonzepte notwendig.

Je höher die Temperatur ist, bei der eine Schaltung funktionieren muss, desto kürzer ist ihre Lebensdauer, da die meisten Ausfallmechanismen durch Temperatur beschleunigt werden. Unter extremen Umgebungsbedingungen bekommen Zuverlässigkeitsaspekte deshalb eine herausragende Bedeutung. Dabei ist allein schon die Vorgehensweise der Untersuchungen an sich eine besondere Herausforderung, da Testsysteme speziell angepasst werden müssen.

Diese Arbeit beschäftigt sich mit Zuverlässigkeitsuntersuchungen an einem Hochtemperatur-CMOS-Prozess auf SOI-Substraten mit einer Strukturgröße von 1 µm, der am Fraunhofer IMS für den Einsatz zwischen -40 °C und 250 °C entwickelt wurde. Um die Lebensdauer und damit die Zuverlässigkeit eines Bauelements oder einer Schaltung bei der maximalen Betriebstemperatur (z. B. 150 °C) abschätzen zu können, werden beschleunigte Zuverlässigkeitsuntersuchungen bei Temperaturen bis 250 °C durchgeführt. In dem hier vorliegenden Fall entspricht die übliche maximale Testtemperatur aber schon der maximalen Betriebstemperatur, weshalb beschleunigte Qualitätsuntersuchungen bei Temperaturen oberhalb von 250 °C nötig werden. Dabei ist es nicht ausgeschlossen, dass bei Testtemperaturen von mehr als 250 °C neue Ausfallmechanismen auftreten, die bei niedrigeren Temperaturen keine Rolle spielen.

Bislang gibt es in der Literatur nur vereinzelte Untersuchungen zur Zuverlässigkeit bzw. zum Betrieb von CMOS-Schaltungen oberhalb von 250 °C. Zudem werden die meisten Untersuchungen an Bauelementen auf Bulk-Substraten durchgeführt, die eigentlich gar nicht für

den Einsatz bei 250 °C gedacht sind, weshalb dort bei hohen Temperaturen möglicherweise andere Phänomene auftreten als bei den hier verwendeten SOI-Substraten. Ein weiterer wichtiger Bestandteil dieser Arbeit sind Untersuchungen zur Langzeitstabilität für eine Dauer von bis zu 10.000 h bei 250 °C, wie zum Beispiel bei den Ringoszillatoren in Betrieb. Häufig sind Zuverlässigkeitsuntersuchungen darauf bedacht, möglichst schnell Aussagen zu liefern, aber erst die tatsächliche Beobachtung von Schaltungen über einen so langen Zeitraum hinweg kann sicher ihre Funktionalität bestätigen (oder widerlegen).

Ein weiterer Aspekt, der sich bei Untersuchungen oberhalb von 250 °C automatisch ergibt, ist die Frage, bis zu welcher Temperatur CMOS-Bauelemente und -Schaltungen auf SOI überhaupt nutzbar sind. Elektronik auf Siliziumbasis kennt ein natürliches Limit der Funktionalität, wenn Silizium intrinsisch wird, d. h. wenn eine große Anzahl thermisch angeregter Ladungsträger den Unterschied zwischen n- und p-dotierten Bereichen aufhebt. Inwieweit Bauelemente und Schaltungen aber bis zu diesem Limit nutzbar sind, wird in dieser Arbeit evaluiert.

Insgesamt bietet diese Arbeit einen umfassenden Überblick über die wichtigsten Zuverlässigkeitsphänomene und Ausfallmechanismen, die CMOS-Bauelemente und CMOS-Schaltungen auf SOI bei Temperaturen von 250 °C und mehr betreffen können.

Die Untersuchungen umfassen dabei die wesentlichen Prozesskomponenten, die auch bei einer Technologiequalifizierung betrachtet werden würden. Dazu gehören das Gateoxid und die Metallisierung, Untersuchungen des Ladungserhalts und der Zyklenfestigkeit von EEPROM-Speicherzellen und die Analyse des Parameterdrifts von Transistoren in Betrieb (NBTI, Hot Carrier). Zudem werden auch Langzeituntersuchungen durchgeführt und die Bauelemente und Testschaltungen bis 450 °C charakterisiert. Soweit es möglich ist, orientiert sich die Vorgehensweise der Messungen an den in der Mikroelektronik für Zuverlässigkeitsanalysen häufig verwendeten JEDEC-Standards.

Die Untersuchungen sind in fünf Kapitel unterteilt:

Kapitel 2 stellt zunächst den Hochtemperatur-SOI-CMOS-Prozess „H10“ vor, an dem die Untersuchungen durchgeführt wurden. Außerdem geht es auf grundlegende Aspekte von Zuverlässigkeitsuntersuchungen, die JEDEC-Standards sowie den Messaufbau ein.

In **Kapitel 3** geht es um die Qualität und die Zuverlässigkeit des Gateoxids bei 250 °C und darüber hinaus. Hauptaspekt der Untersuchungen sind dabei Messungen zum zeitabhängigen dielektrischen Durchbruch (TDDB).

Kapitel 4 setzt sich mit der Zuverlässigkeit von EEPROM-Speicherzellen auseinander. In Bezug auf den Datenerhalt werden die bei Lagerungstemperaturen oberhalb von 250 °C in Erscheinung tretenden Aspekte herausgearbeitet. Zudem wird die Abhängigkeit der Zyklenfestigkeit von verschiedenen Parametern untersucht.

In **Kapitel 5** liegt der Fokus der Untersuchungen auf der Zuverlässigkeit der Metallisierung, die anhand von Elektromigrationstests und Stressmigrationstests analysiert wird.

Kapitel 6 umfasst Zuverlässigkeitsuntersuchungen und Charakterisierungen von Transistoren und einfachen Grundschaltungen bis 450 °C. Neben den Phänomenen der Parameterinstabilität bei negativer Gate-Source-Spannung (NBTI) und der Auswirkung „heißer“ Ladungsträger

(Hot Carrier) wird die Langzeitstabilität von Ringoszillatoren und Bandgap-Schaltungen bei 250 °C (und soweit möglich auch bei 350 °C) betrachtet.

Kapitel 2

Hochtemperaturelektronik und Zuverlässigkeit

In diesem Kapitel sollen die beiden Themenkomplexe dieser Arbeit, die Hochtemperaturelektronik und die Zuverlässigkeit, genauer erläutert werden. Zunächst wird der Hochtemperatur-SOI-CMOS-Prozess „H10“ des Fraunhofer IMS, an dem die Zuverlässigkeitsuntersuchungen durchgeführt wurden, genauer beschrieben. Dann geht ein weiterer Unterabschnitt auf grundlegende Konzepte für die Messung und Bewertung der Zuverlässigkeit von Halbleitern ein. Ein besonderes Augenmerk wird dabei auf die JEDEC-Standards gerichtet. Der letzte Abschnitt dieses Kapitels stellt den Messaufbau für die Untersuchungen vor.

2.1 Der Hochtemperatur-SOI-CMOS-Prozess „H10“

Die Zuverlässigkeitsuntersuchungen in dieser Arbeit wurden an Bauelementen vorgenommen, die im Hochtemperaturprozess „H10“ des Fraunhofer IMS gefertigt wurden. Dieser Prozess basiert auf Dünnfilm-Silicon-on-Insulator-Wafern mit einem Durchmesser von 200 mm und ist für die Anwendung bei Temperaturen zwischen -40 °C und 250 °C vorgesehen. Abbildung 2.1 zeigt einen schematischen Querschnitt durch den Wafer.

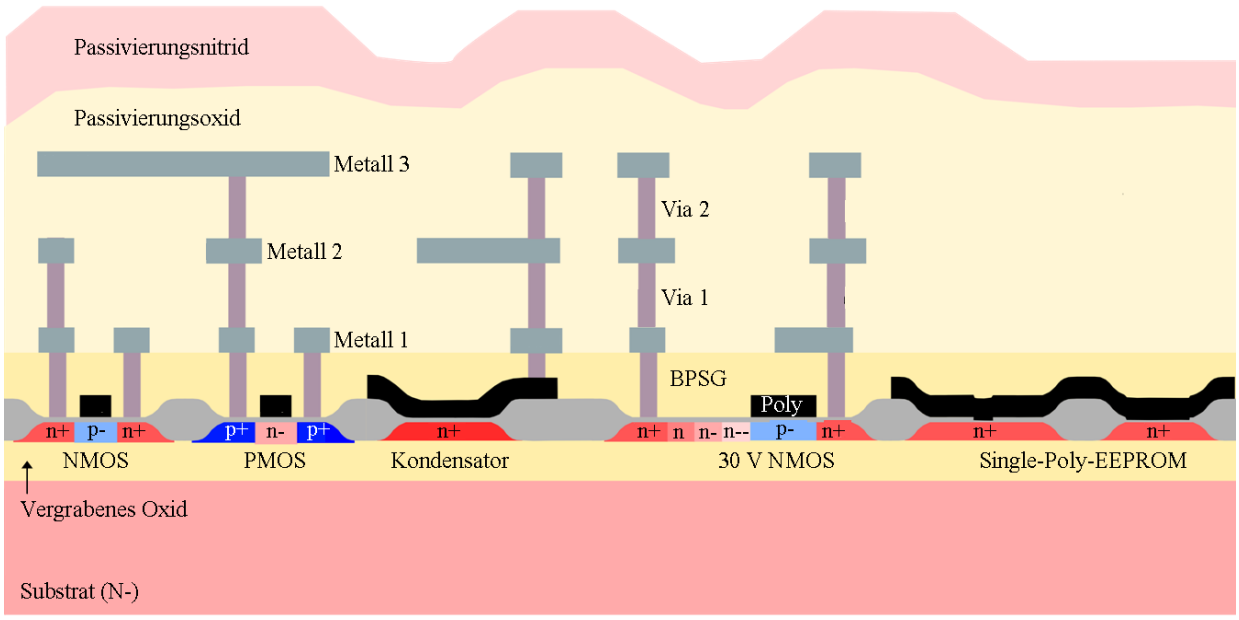


Abbildung 2.1: Schematischer Querschnitt eines H10-Wafers mit einigen der möglichen Bauelemente; aus [DMH10]

Mit einer Gateoxiddicke von 40 nm, einer Tunneloxiddicke von 11 nm und drei Lagen einer Wolframmetallisierung[1] ergeben sich zahlreiche Optionen für digitale und analoge Schaltungen. Dazu gehören Polysilizium-Nplus-Kondensatoren, Zener- und Pindioden, Single-Poly-EEPROM-Zellen und 30 V - Transistoren. Für digitale Anwendungen bei 3,3 V stehen Transistoren mit einer Kanallänge von 1 µm, für analoge Anwendungen bei 5 V Transistoren mit einer Kanallänge von 1,6 µm zur Verfügung. Die NMOS-Transistoren haben eine Schwellenspannung von etwa 1 V bei Raumtemperatur. Daneben sind zwei verschiedene PMOS-Transistoren vorhanden. Bei Raumtemperatur besitzt der eine PMOS-Transistor eine Schwellenspannung von etwa -1 V und der andere PMOS-Transistor eine Schwellenspannung von etwa -2 V.

Der entscheidende Grund für die Verwendung von Silicon-on-Insulator, kurz SOI genannt, sind die geringeren Leckströme im Vergleich zur konventionellen Bulk-Technologie. Abbildung 2.2 demonstriert den Unterschied zwischen einem NMOS-Transistor in Bulk- und einem NMOS-Transistor in SOI-Technologie. Auf Bulk-Wafern sind die Bauelemente direkt in den Siliziumwafer integriert, wodurch pn-Leckströme, die mit der Temperatur stark zunehmen, unvermeidbar sind. Eine sinnvolle Schaltungsfunktion kann deshalb in der Bulk-CMOS-Technologie nur bis Temperaturen von maximal 175 °C gewährleistet werden. Bei der SOI-Technologie ist ein dünner Siliziumfilm (im H10-Prozess ca. 150 nm dick) durch eine Oxidschicht, dem vergrabenen Oxid (engl. Buried Oxide, kurz BOX), vom Siliziumsubstrat getrennt. Die Bauelemente werden in dem dünnen Siliziumfilm realisiert und sind damit dielektrisch vom Substrat isoliert. Im Gegensatz zur Bulk-Technologie haben Bauelemente auf SOI-Wafern bei höheren Temperaturen dann ein deutlich geringeres Leckstromniveau. Die SOI-CMOS-Technologie ist deshalb für Hochtemperaturanwendungen gut geeignet.

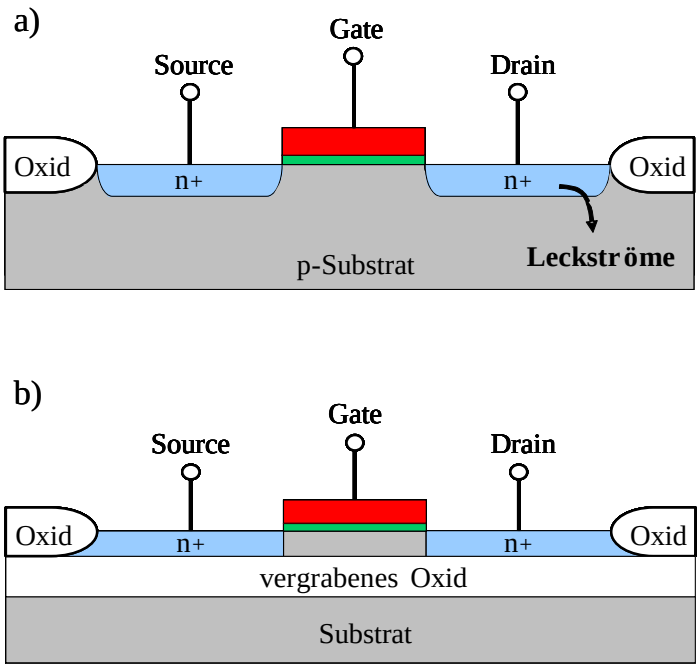


Abbildung 2.2: a) Schematischer Querschnitt eines NMOS-Transistors in Bulk-Technologie; b) Schematischer Querschnitt eines NMOS-Transistors in SOI-Technologie

[1] Zu Testzwecken wurden auch Wafer statt mit einer Wolframmetallisierung mit ein bis drei Lagen Aluminium versehen. Diese Wafer finden zum Beispiel bei den Untersuchungen an EEPROM-Zellen Verwendung, wo die Metallisierung an sich nicht Aspekt der Untersuchung ist.

Neben dem beschriebenen H10-Prozess wird am Fraunhofer IMS noch ein weiterer Hochtemperatur-SOI-CMOS-Prozess entwickelt, der ebenso für Betriebstemperaturen von 250 °C gedacht ist. Der „H035“ genannte Prozess hat eine minimale Strukturgröße von 0,35 µm, zwei Gateoxiddicken von 10 nm und 40 nm und besitzt eine vier-lagige Wolframmetallisierung sowie eine EEPROM-Prozessoption. Da sich dieser Prozess noch in der frühen Entwicklung befindet, wurden die Zuverlässigkeitsuntersuchungen ausschließlich an H10-Wafern durchgeführt.

2.2 Zuverlässigkeit in der Mikroelektronik

Unter dem Begriff „Zuverlässigkeit“ versteht man die Fähigkeit, eine bestimmte Funktion in einer bestimmten Umgebung während einer definierten Anwendungsdauer zu erfüllen. Ist ein Produkt zuverlässig, bedeutet dies, dass es im vorgesehenen Rahmen einwandfrei funktioniert und nutzbar ist.

Die Zuverlässigkeit von Bauelementen wird durch die Anzahl der Elemente bestimmt, die zu einem bestimmten Zeitpunkt ausgefallen sind, d. h. die die definierten Anforderungen nicht mehr erfüllen. Bei der zeitlichen Verteilung dieser Ausfälle unterscheidet man drei wesentliche Abschnitte: die Phase der sogenannten „infant mortality“, den Ausfall während der Betriebsphase und den Ausfall durch Alterungserscheinungen (“wearout“). Abbildung 2.3 zeigt diese drei Abschnitte in einer Kurve, die aufgrund ihrer Form als „Badewannenkurve“ bezeichnet wird.

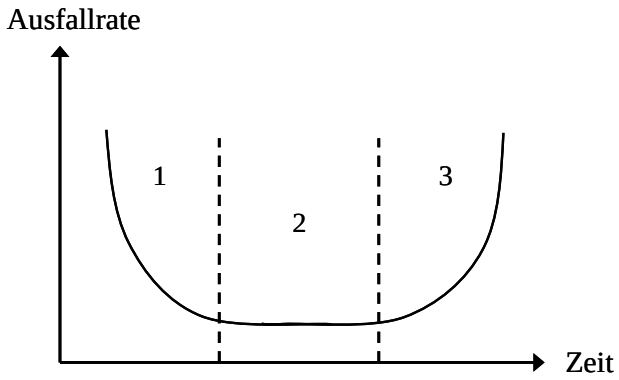


Abbildung 2.3: „Badewannenkurve“: Ausfallverteilung von Bauelementen mit drei Abschnitten 1. „infant mortality“, 2. Betriebsphase, 3. Alterung („wearout“)

Für den ersten Teil der Kurve sind Partikel, Unreinheiten oder Strukturdefekte während der Prozessierung verantwortlich. Ziel ist ein steiler erster Kurvenabschnitt, um schlechte Stücke schnell aussortieren zu können. Der Grund für Ausfälle während des zweiten Abschnitts sind zufällige Fehler. Diese zweite Phase sollte möglichst lange sein, d. h. der Beginn des dritten Abschnitts sollte möglichst spät erfolgen. Die Zuverlässigkeitsanalyse, die auch Gegenstand dieser Arbeit ist, beschäftigt sich dann mit dem dritten Kurvenabschnitt, d. h. mit der Alterung von Bauelementen.

Die Zuverlässigkeitsuntersuchungen in dieser Arbeit werden an fertig prozessierten Bauelementen wie Kondensatoren, Transistoren und Schaltungen oder auch speziellen Teststrukturen wie z.B. bei der Untersuchung der Metallisierung vorgenommen. Mechanische Tests

und die Untersuchung der Aufbau- und Verbindungstechnik sind kein Gegenstand dieser Arbeit.

Um zu untersuchen, wie sich die Zuverlässigkeit an sich und die Messung und Analyse der Zuverlässigkeit bei hohen Temperaturen von denen bei niedrigeren Temperaturen unterscheiden, muss die Grundlage der Hochtemperaturuntersuchungen mit der prinzipiellen Vorgehensweise bei niedrigeren Temperaturen vergleichbar sein. Deshalb werden in dieser Arbeit, so weit es möglich ist, die Methoden der JEDEC-Standards genutzt. Ergeben sich bei hohen Testtemperaturen oder aufgrund sonstiger Einflüsse Abweichungen, wird dies entsprechend diskutiert und analysiert. Die Tests werden an Teststrukturen durchgeführt, die konform zu den JEDEC-Standards entworfen wurden.

2.3 JEDEC-Standards und grundlegende Konzepte für Zuverlässigkeitsanalysen

Die Zuverlässigkeit eines Produktes ist ein wichtiger Aspekt bei Entwicklung, Produktion und Vertrieb. Deshalb ist es sowohl aus Sicht des Herstellers als auch aus Sicht des Käufers von Vorteil, wenn bei der Beurteilung der Zuverlässigkeit eines Produktes allgemein anerkannte Standards eingehalten werden. Der Käufer kann nachvollziehen, nach welchen Kriterien das Produkt getestet wurde, und der Verkäufer kann seine eigenen Produkte mit anderen besser vergleichen und belegen, welchen Bedingungen das Produkt standhält und welchen nicht. Solche Standards existieren für verschiedene Anwendungsbereiche, zum Beispiel die MIL-Standards für militärische Applikationen, AEC-Standards für den Automobilbereich oder die hier verwendeten JEDEC-Standards für die Halbleiterindustrie.

Die JEDEC-Standards sind Übereinkünfte oder Vereinheitlichungen von Methoden zur Qualifizierung und Zuverlässigkeitsanalyse für die Halbleitertechnik. Sie sind von Komitees der US-amerikanischen JEDEC Solid State Technology Association erarbeitet worden, deren Mitglieder Qualitätsingenieure von Halbleiterfirmen sind. Im Gegensatz zu Normen wie den DIN- oder ISO-Normen haben die JEDEC-Standards einen weniger verbindlichen Charakter. Sie sind aber in der Halbleiterindustrie soweit anerkannt, dass sie de facto nicht umgangen werden können. Viele der JEDEC-Standards werden regelmäßig überarbeitet, wenn neue wissenschaftliche Erkenntnisse oder veränderte Ansprüche eine Aktualisierung nötig machen.

Das JEDEC-Dokument JP001 [JED1] ist eine häufig beachtete Zusammenstellung verschiedener Zuverlässigkeitsaspekte, das Informationen über Testmethoden, Teststrukturen und Analyseverfahren enthält. Aufgrund unterschiedlicher Interessen der verschiedenen Halbleiterfirmen ist es kein Standard, sondern „nur“ eine Veröffentlichung von Vorschlägen. Dennoch beziehen sich viele Zuverlässigkeitsanalysen auf dieses Dokument. Für die in dieser Arbeit durchgeführten Messungen fand das JEDEC-Dokument JP001 häufig Verwendung. Die Untersuchungen wurden aber auch auf Basis weiterer JEDEC-Dokumente durchgeführt. Tabelle 2.1 gibt einen Überblick über die verwendeten Standards (JESD) und Publikationen (JP, JEP)[2]. Die genaue Vorgehensweise der in den einzelnen Dokumenten beschriebenen Prozeduren wird in den jeweiligen Kapiteln genauer erläutert.

Bei Zuverlässigkeitsuntersuchungen muss man zwischen dem Zuverlässigkeitsaspekt an sich (engl. failure mode) und dem dahinter stehenden physikalischen Mechanismus (engl. failure mechanism) unterscheiden. Für den Themenbereich der EEPROM-Zellen gibt es beispiels-

[2] Die aufgelisteten Standards und Dokumente sind unter www.JEDEC.org zu finden.

weise den Aspekt des Datenerhalts bzw. des Datenverlusts. Der Datenverlust kann aber viele verschiedene Ursachen haben, wie zum Beispiel Oxiddefekte, dielektrische Durchbrüche des Tunneloxids oder Kontaminationen. Jeder dieser physikalischen Mechanismen tritt unter anderen Bedingungen (Strom, Spannung, Temperatur, Prozessierung etc.) auf und hat eine andere Aktivierungsenergie. Bei Zuverlässigkeitsuntersuchungen geht es deshalb darum, für die relevanten Themenbereiche verschiedene Zuverlässigkeitsaspekte zu untersuchen und herauszufinden, welche physikalischen Mechanismen für Degradationen bzw. Ausfälle der Bauelemente verantwortlich sind.

Themenbereich	**Zuverlässigkeitsaspekt**	**JEDEC-Dokument**
Gateoxid	Zeitabhängiger dielektrischer Durchbruch (TDDB)	JP001, JESD92
EEPROMs	Datenerhalt	JESD22-A117, JESD22-A103C
	Zyklenfestigkeit	JESD22-A117
Metallisierung	Elektromigration	JP001, JESD61, JESD63, JESD87
	Stressmigration	JP001, JESD87, JEP139
Transistoren	Hot Carrier (HCI)	JP001, JESD28-A, JESD28-1, JESD60A
	Negative Bias Temperature Instability (NBTI)	JP001, JESD90
Allgemein	Fehlermechanismen, Modelle und Aktivierungsenergien	JEP122E

Tabelle 2.1: Übersicht über die in der vorliegenden Arbeit verwendeten JEDEC-Dokumente

Ein weiterer wichtiger Punkt ist die Unterscheidung zwischen Aussagen über die Qualität eines Bauelements und seiner Lebensdauer. Es gibt einige Messmethoden (z. B. QBD, siehe Kapitel 3), die nur dem Vergleich verschiedener Wafer oder Chargen dienen und damit Aussagen über die Qualität, nicht aber über die Lebensdauer eines Bauelements machen können, weil bei der Untersuchung keine Ausfallzeiten gemessen werden. Bei allen in dieser Arbeit verwendeten Methoden ist die Messung einer Ausfallzeit vorgesehen, so dass eine Abschätzung der Lebensdauer der Bauelemente prinzipiell möglich ist.

Fast alle der hier untersuchten Zuverlässigkeitsaspekte (bis auf die Auswirkungen von Hot Carrier - Stress) haben mit zunehmender Temperatur stärkere Auswirkungen, d. h. sowohl das Gateoxid, die Metallisierung, der Datenerhalt und die Zyklenfestigkeit von EEPROM-Zellen als auch die Funktionalität von Transistoren degradieren mit steigender Temperatur. Aus diesem Grund können zur Beurteilung der Zuverlässigkeit bei Betriebstemperatur Messungen bei höheren Temperaturen durchgeführt und damit der Ausfall des Bauelements schneller herbeigeführt werden, wenn der Zusammenhang zwischen der Test- und der Betriebstemperatur bekannt ist. In den meisten Fällen ist eine Extrapolation von der Test- zur Betriebstemperatur über das Arrheniusgesetz oder über abgewandelte Formen dieses Gesetzes möglich. Im Allgemeinen beschreibt das Arrheniusgesetz die Temperaturabhängigkeit der Reaktionsgeschwindigkeit eines Prozesses, d. h. den Zusammenhang zwischen der Zeit der Reaktion und der Temperatur [ARR89]. Für die Lebensdauer eines Bauelements bei einer bestimmten Temperatur und bei Vorliegen eines bestimmten Ausfallmechanismus ergibt sich dann folgende Gleichung:

$$t_{\text{fail}} = t_0 \cdot \exp\left(\frac{E_A}{k \cdot T}\right) \qquad [2.1]$$

Dabei ist t_{fail} die Zeit bis zum Ausfall des Bauelements, t_0 eine Zeitkonstante, E_A die Aktivierungsenergie, k die Boltzmann-Konstante und T die Temperatur. Die Aktivierungsenergie ist die Energie, die aufgebracht werden muss, um einen bestimmten Mechanismus auszulösen. Sie ist damit für jeden Ausfallmechanismus spezifisch. Kennt man die Ausfallzeit bei der Testtemperatur und die Aktivierungsenergie für den Mechanismus, kann die Ausfallzeit bei der Betriebstemperatur ermittelt werden. In der Praxis wird entweder eine Aktivierungsenergie aus der Literatur für einen bekannten physikalischen Mechanismus angenommen oder über verschiedene Testtemperatur-Ausfallzeit-Kombinationen selbst ermittelt.

Eine weitere Möglichkeit, die Zuverlässigkeitsanalysen zu beschleunigen und damit den Ausfall des Bauelements schneller herbei zu führen, ist das Stressen durch erhöhten Strom oder erhöhte Spannung. Hierbei gibt es für die verschiedenen Ausfallmechanismen jeweils spezifische Modelle, um auf die Betriebsbedingungen extrapolieren zu können. In den meisten Fällen werden Temperaturbeschleunigung und Strom- oder Spannungsbeschleunigung kombiniert.

Die JEDEC-Dokumente geben an, wie der Ausfall eines Bauelementes charakterisiert wird. Beispielsweise werden für die Ermittlung der Lebensdauer des Gateoxids Kondensatoren so lange gestresst, bis Strom ungehindert durch das Oxid fließen kann und die Isolatoreigenschaften unwiederbringlich verloren sind. Die Ausfallzeit wird dann mit t_{BD}, also mit der „Zeit bis zum Durchbruch“ (engl. time to breakdown) bezeichnet. Im Falle der Metallisierung wird hingegen nicht notwendigerweise bis zum Totalausfall gestresst. Hier kennzeichnet eine festgelegte Widerstandserhöhung den Ausfall der Metallbahn, des Vias oder des Kontakts. Die Ausfallzeit ist dann mit t_{fail} die „Zeit bis zum Ausfall“ (engl. time to failure). Auch bei Hot Carrier- oder NBTI-Untersuchungen geht es nicht um einen Totalausfall des Bauelements, sondern um das Erreichen einer zuvor festgelegten Degradation. t_{tar} bezeichnet dann die „Zeit bis zum Erreichen einer festgelegten Änderung“ (engl. time to target).

Im Rahmen dieser Arbeit wird die Zuverlässigkeit des H10-Prozesses bei Temperaturen von 250 °C und mehr analysiert. In den meisten JEDEC-Dokumenten werden aber nur Tests bei der maximalen Betriebstemperatur (z.B. 150 °C) oder zur Beschleunigung bei Temperaturen bis maximal 200 °C empfohlen (z.B. für die TDDB-Messungen). Für Temperaturen von 250 °C und mehr können aber andere Fehlermechanismen auftreten, so dass die in den JEDEC-Standards beschriebenen Messmethoden und Modelle möglicherweise nicht mehr nutzbar sind. Ein Aspekt in dieser Arbeit ist deshalb die Überprüfung der Anwendbarkeit der in den JEDEC-Dokumenten beschriebenen Prozeduren bei Temperaturen von 250 °C und mehr.

2.4 Messaufbau für die Zuverlässigkeitsuntersuchungen

Die Untersuchungen zur Qualität und Zuverlässigkeit des H10-Prozesses wurden zum Teil auf Waferebene an manuellen Waferprobern, zum Teil an Bauelementen im Gehäuse in zwei verschiedenen Öfen durchgeführt. Die Maximaltemperatur für Untersuchungen auf Waferebene beträgt 300 °C. Um den Temperaturbereich bis 450 °C zu erweitern und auch Langzeituntersuchungen und Waferlagerungen zu ermöglichen, wurden eigens für diese Untersuchungen zwei Öfen angeschafft. In den folgenden Unterkapiteln werden die beiden Messsysteme vorgestellt.

Zum Anlegen und Messen von Strom oder Spannung wurden Semiconductor Parameter-analyzer und Pulsgeneratoren von Hewlett Packard bzw. Agilent verwendet[3]. Die Messgeräte wurden mit am Institut erarbeiteten LabVIEW-Programmen[4] oder mit der Software ICS[5] angesteuert.

2.4.1 Untersuchungen auf Wafer-Ebene

Eine geeignete Methode, um eine größere Anzahl an einzelnen Bauelementen wie Kondensatoren oder Transistoren elektrisch zu testen, ist die Untersuchung auf Waferebene. Dabei wird das Bauelement direkt auf dem Wafer kontaktiert. Die JEDEC-Organisation empfiehlt in ihren Teststandards in den meisten Fällen den Test auf Waferebene (engl. wafer level). Deshalb spricht man auch von Wafer Level Reliability, kurz WLR.

Abbildung 2.4 zeigt ein Foto eines Waferprobers, auch Spitzenmessplatz genannt. Der Wafer (8'' bzw. 200 mm Durchmesser) wird auf dem Thermochuck (a) platziert und mit Vakuum festgehalten. Der Chuckaufsatz, auf dem der Wafer liegt, ist elektrisch vom restlichen System getrennt und kann mit einem Potential versehen werden. Für die Messungen von SOI-Wafern ist es somit möglich, das Back-Gate separat zu kontaktieren. Der Chuck kann außerdem über eine externe Steuerung auf bis zu 300 °C geheizt werden. Die einzelnen Testfelder auf dem Wafer sind mit Hilfe des Mikroskops (b) sichtbar. Der Chuck lässt sich sowohl horizontal in *x*- und *y*-Richtung (c) als auch vertikal in *z*-Richtung (d) bewegen. Mit Messspitzen bzw. Messnadeln aus Wolfram (e), deren Spitzendurchmesser 2 µm beträgt, können die Pads, an die die Messstrukturen angeschlossen sind, kontaktiert werden. Die Messnadeln sind an den Armen von Manipulatoren (f) befestigt. Für Messungen bis 300 °C stehen spezielle hitzebeständige Manipulatoren zur Verfügung. Die Manipulatoren werden durch Vakuum auf einer Plattform fixiert (g). Triaxialkabel (h) führen von den Manipulatoren zu den Mess-geräten.

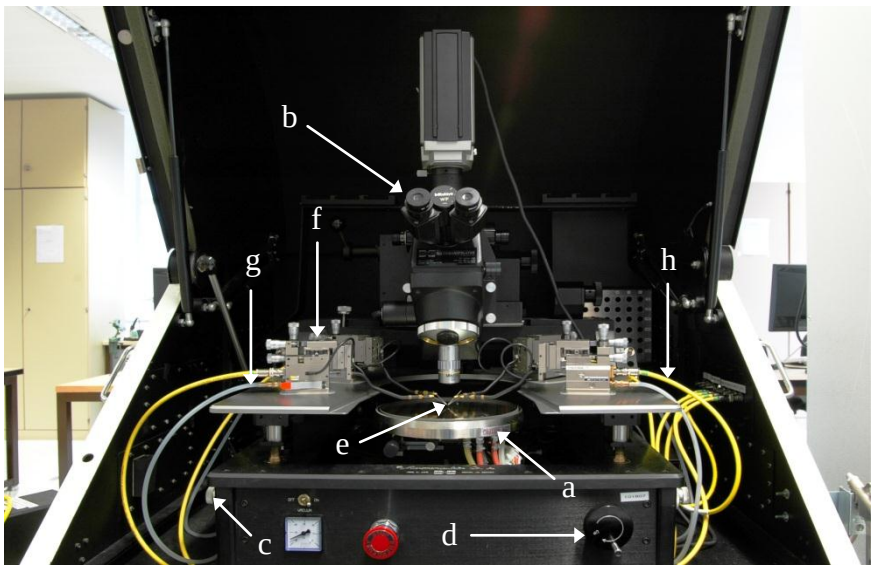


Abbildung 2.4: Foto eines Waferprobers mit den einzelnen Komponenten: a) Thermochuck, b) Mikroskop, c) *x*- und *y*-Bewegung, d) *z*-Bewegung, e) Messnadeln, f) Manipulatoren, g) Vakuumzufuhr, h) Triaxialkabel

[3] Weitere Informationen dazu sind unter www.agilent.com zu finden.
[4] LabVIEW ist eine graphische Programmiersprache von National Instruments (www.ni.com).
[5] ICS ist eine Messsoftware der Firma Metrics (www.metricstech.com).

Am IMS stehen zwei solcher Waferprober für einen Temperaturbereich von Raumtemperatur bis 300 °C zur Verfügung. Ein weiterer Spitzenmessplatz besitzt einen Thermochuck, der bis -40 °C gekühlt und bis 125 °C geheizt werden kann. Dieser Waferprober fand bei den Untersuchungen zum Hot Carrier - Stress Anwendung.

Der wesentliche Vorteil von Messungen auf Waferebene ist das einfache Verfahren der Messnadeln von einem Testfeld zum Nächsten. Benötigt man für statistische Messungen mehrere identische Strukturen, kann man leicht eine andere Teststruktur kontaktieren, da man den ganzen Wafer auf einmal zur Verfügung hat. Nachteilig ist aber die Beschränkung auf Temperaturen bis 300 °C.

2.4.2 Untersuchungen an Bauelementen im Gehäuse und Langzeitlagerungen

Für den erweiterten Temperaturbereich stehen zwei Öfen zur Verfügung. Der eine Ofen ist für Langzeituntersuchungen und Waferlagerungen bei 250 °C vorgesehen, im anderen Ofen werden kürzer terminierte Versuche bis 450 °C durchgeführt. Beide Öfen sind mit einer Stickstoffzuleitung und einer Abluftleitung versehen, jedoch nicht hermetisch dicht. An den seitlichen Außenwänden sind Durchführungen angebracht, durch die spezielle Hochtemperaturkabel vom Inneren des Ofens nach außen geführt werden können. Eine Abdichtung dieser Zuleitungen um die Kabel herum gewährleistet, dass die Temperatur im Ofen stabil bleibt. Beide Öfen besitzen eine forcierte Umluftzirkulation für eine gute räumliche Temperaturverteilung.

Für die Messungen werden gesägte Chips mit Silberglaslot[6] in Keramikgehäuse geklebt und die Chippads mit den Gehäusepads durch Aluminiumdraht verbunden. Dafür haben die Chippads alle eine zusätzliche Aluminiumlage auf der Wolframmetallisierung. Diese Aluminiumlage ist auch bei Untersuchungen auf Waferebene von Vorteil, da Messnadelabdrücke darauf leichter zu erkennen sind als bei direktem Kontakt auf Wolfram und der Kontakt zwischen einer Nadel und einem Aluminiumpad stabiler ist als zwischen einer Nadel und einem Wolframpad (siehe auch Kapitel 5). Die Keramikgehäuse werden dann in Keramiksockel gesteckt, an die wiederum Hochtemperaturkabel befestigt sind (siehe Abbildung 2.5 b). Die Sockel werden im Ofen an einer speziell konstruierten Halterung platziert und die Kabel nach außen geführt. Außerhalb des Ofens sind die Verbindung auf Triaxialkabel und der Anschluss an die Messgeräte mit entsprechenden Steckverbindern möglich. In Abbildung 2.5 a) ist das Innere eines Ofens mit der Halterung für die Sockel und den Hochtemperaturleitungen zu sehen.

Im Vergleich zu den Messungen auf Waferebene haben die Untersuchungen an aufgebauten Bauelementen einige Nachteile. Im Gegensatz zur Waferebene, wo leicht mehrere Strukturen nacheinander vermessen werden können, muss hier für jede untersuchte Struktur ein anderer Chip angefertigt und im Ofen platziert werden. Zudem sind die Messungen auf einen guten Kontakt zwischen Chippad und Gehäusepad, zwischen Gehäusepad und Sockel sowie zwischen Sockel und Hochtemperaturleitung angewiesen. Dieser Kontakt muss auch bei 450 °C stabil sein. Da zum Zeitpunkt der Messungen eine Aufbau- und Verbindungstechnik nur für Aluminiumpads mit Aluminiumdraht zur Verfügung stand und zudem bei Temperaturen von 450 °C schon nach wenigen Stunden die Wolframmetallisierung unter den Aluminiumpads zu oxidieren beginnt (vergleiche Kapitel 5), wurden Langzeittests nur bis 350 °C durchgeführt. Die Messdauer bei 450 °C wurde auf maximal eine Stunde beschränkt,

[6] „Leitfähiger Silber/Glas-Kleber DAP 1“ von Ferro Electronic Material Systems (www.ferro.com).

um zwischen dem Ausfall des Bauelements und der Degradation durch den Aufbau zu unterscheiden. Wenn die Verbindung zwischen dem Bauelement und dem Messgerät zuverlässig hergestellt ist, und die zeitliche Einschränkung für höhere Temperaturen beachtet wird, ist der Aufbau aber stabil.

a) b)

Abbildung 2.5: a) Ofen mit der Halterung für die Sockel mit Keramikgehäusen und nach außen geführten Hochtemperaturkabeln; b) Sockel mit Keramikgehäuse und aufgeklebtem Chip

Ein großer Vorteil der Öfen ist neben der Tatsache, dass bei höheren Temperaturen als auf dem Waferprober gemessen werden kann, die Möglichkeit, die Öfen parallel für Messungen und Lagerungen zu nutzen.

Kapitel 3

Qualität und Zuverlässigkeit des Gateoxids

Dieses Kapitel beschäftigt sich mit der Qualität und der Zuverlässigkeit des im untersuchten Hochtemperaturprozess verwendeten Gateoxids. Während im Zusammenhang mit der fortschreitenden Miniaturisierung in der Mikroelektronik die Gateoxiddicke immer weiter reduziert wird, um die Geschwindigkeit von Schaltungen zu steigern, wird im Falle hoher Einsatztemperaturen eine ausreichend große Oxiddicke benötigt, um die Spannungsfestigkeit der Bauelemente und Schaltungen zu gewährleisten. Das Gateoxid des H10-Prozesses besteht deshalb aus einer Schicht von 40 nm thermisch gewachsenem Siliziumdioxid (SiO_2) und wird in Kondensatoren und Transistoren zur Isolation des Polysilizium-Gates vom Siliziumfilm verwendet.

Wie in jeder Technologie gibt es auch hier weitere Oxide wie das vergrabene Oxid, das Feldoxid, Zwischenoxide zwischen den Metalllagen oder das Oxid zur Passivierung. Alle diese Oxide sind wesentlich dicker und elektrisch weniger belastet als das Gateoxid und deshalb von Zuverlässigkeitsproblemen weniger stark betroffen. Sie werden in dieser Arbeit nicht behandelt. Das Tunneloxid wird für EEPROM-Speicherzellen verwendet (siehe Kapitel 4). Ist in diesem Kapitel von „Oxid“ die Rede, handelt es sich immer um das Gateoxid.

Bei der Untersuchung der Oxid-Zuverlässigkeit sind im Allgemeinen zwei Aspekte von Bedeutung: Schäden durch elektrischen Stress und Strahlenschäden. Da hohe Umgebungstemperaturen vor allem die elektrischen Eigenschaften des Oxids beeinflussen und Strahlungsbelastung in diesem Zusammenhang keine Rolle spielt, beziehen sich die vorgestellten Untersuchungen deshalb ausschließlich auf die Ursachen und Eigenschaften des elektrischen Oxiddurchbruchs bei hohen Temperaturen.

3.1 Theoretische Aspekte zur Zuverlässigkeit des Gateoxids

3.1.1 Ladungen und Haftstellen im Oxid

Die Belastung durch Strom und Spannung kann eine Oxidschicht verändern und dauerhaft schädigen. Wie es zu dieser Schädigung kommt, wird in verschiedenen Modellen unterschiedlich begründet. Alle Modelle gehen jedoch davon aus, dass im Oxid Ladungen vorhanden sind, die auf unterschiedliche Weise den Stromfluss durch das Oxid beeinflussen.

Die Oxidladungen werden in vier Gruppen unterteilt (siehe auch Abbildung 3.1), die nach ihrer räumlichen Lage und ihrer Fähigkeit, mit anderen Ladungen in Kontakt zu treten, unterschieden werden [DEA80], [SCH98], [SOM10].

1.) Feste Ladungen (engl. fixed charges)

2.) Bewegliche Ladungen (engl. mobile charges)

3.) Grenzflächenladungen oder Grenzflächenhaftstellen (engl. interface trapped charges oder interface traps)

4.) Volumenladungen oder Volumenhaftstellen (engl. bulk oxide trapped charges oder bulk oxide traps)

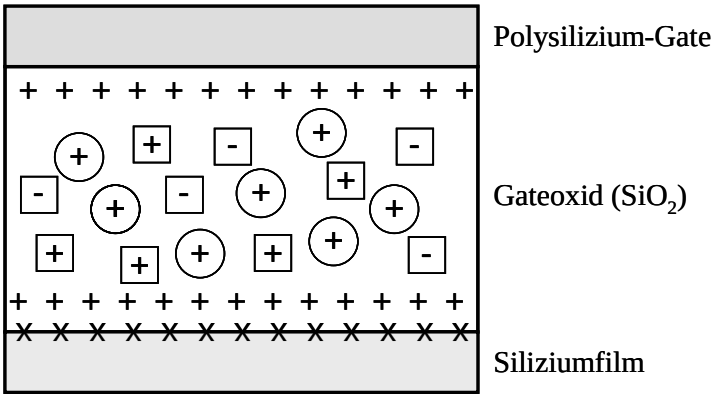


Abbildung 3.1: Ladungen im Gateoxid und an der Grenzfläche zum Silizium: Feste Ladungen (+); bewegliche Ladungen (+, eingekreist); Grenzflächenladungen (X) und Volumenladungen (+ und -, eckig eingefasst); nach [DEA80]

Feste Ladungen sind Strukturdefekte in der Oxidschicht, die während des Herstellungsprozesses entstehen, positiv geladen sind und das elektrische Feld zwischen dem Polysilizium-Gate und dem Siliziumfilm beeinflussen können. Sie befinden sich an den inneren Grenzflächen des Oxids, können aber keine Ladung abgeben oder aufnehmen und sind damit „fest".

Die zweite Gruppe, die **beweglichen Ladungen**, sind mobile Alkaliionen wie Natrium (Na^+) oder Kalium (K^+) und entstehen ebenfalls während der Herstellung (Kontamination). Sie können sich frei im Gateoxid bewegen und sind deshalb während der Umpolung eines elektrischen Feldes, das über dem Gateoxid anliegt, detektierbar (siehe dazu auch die TVS-Messungen in Kapitel 4, Abschnitt 4.3.1).

Die beiden in der obigen Liste letztgenannten Ladungstypen haben die Fähigkeit, andere Ladungen aufzunehmen. Sie werden deshalb auch als Defekt oder Haftstelle bezeichnet, in dem bzw. in der sich Ladung festsetzen kann.

Grenzflächenhaftstellen befinden sich an der Grenzfläche des Gateoxids zum Siliziumfilm. Als Haftstelle (engl. trap) sind sie entweder positiv oder negativ geladen und können somit entweder negative oder positive Ladungen aufnehmen und meistens auch wieder abgeben. Sie werden alternativ auch als Grenzflächenladungen bezeichnet. Eine direkte Wechselwirkung mit dem Kanal ist möglich. Solche Grenzflächenhaftstellen entstehen durch elektrischen Stress, Strahlung, Strukturdefekte oder auch bei der Herstellung und befinden sich in einem Bereich von 2-3 nm Abstand zur Siliziumgrenzfläche. Sind sie weiter entfernt, handelt es sich um Volumenhaftstellen.

Volumenhaftstellen oder Volumenladungen können im gesamten Volumen des Gateoxids vorhanden sein. Sie sind ebenfalls entweder positiv oder negativ geladen und haben ihren Ursprung in der Besetzung eines Volumendefekts durch ein Elektron oder ein Loch. Die Entstehungsursache dieser Defekte ist noch in der Diskussion. Möglich sind sowohl elektrischer Stress oder Strahlung als auch herstellungsprozessbedingte Verunreinigungen oder strukturelle Defekte.

Grenzflächenhaftstellen und Volumenhaftstellen werden im Englischen meistens „trap", teilweise aber auch „state" genannt. Im Deutschen wird neben „Haftstelle" auch die Bezeichnung „Defekt" verwendet. „Defekt" ist ein kristallographischer Begriff und beschreibt eine Unregelmäßigkeit im Kristallgitter des Oxids. Das Wort „Haftstelle" betont die Fähigkeit, Ladung aufzunehmen bzw. abzugeben [FLE93].

In der Literatur werden die Begriffe trap, state, Haftstelle, Defekt oder auch Ladung oft vermischt. Im Wesentlichen kommt es aber nur darauf an, ob es sich um eine positive oder negative Stelle handelt und wo sich diese im Gateoxidbereich befindet. Im Folgenden werden vor allem die Bezeichnungen Haftstelle und Ladung verwendet.

Einige Studien haben gezeigt, dass Haftstellen bei ihrer Entstehung meistens neutral sind und erst geladen werden, wenn sie sich in der Nähe der Elektroden befinden. Haftstellen in Nähe der Anode (+) sind positiv geladen, Haftstellen nahe der Kathode (-) negativ. Da die exakte Verteilung der Ladungen im Oxid bisher unbekannt ist, wird oft eine symmetrische Verteilung in Bezug auf Anode und Kathode angenommen [DUM01]. In der Summe hat man im Allgemeinen eine positive Gesamtladung im Oxid.

Die Anwesenheit von Haftstellen bzw. Ladungen im Gateoxid führt zu einer Erhöhung des Leckstroms durch das Oxid [DUM01] und damit letztlich auch zum Verlust der Isolatoreigenschaften. Umgekehrt entstehen Haftstellen aber auch durch Anwesenheit eines Stromflusses durch das Oxid. In Abschnitt 3.1.2 wird genauer auf die Leitungsmechanismen im Gateoxid eingegangen.

3.1.2 Leitungsmechanismen im Gateoxid

Die Leitungsmechanismen im Gateoxid sind vielfältig und hängen von mehreren Faktoren wie dem Material selbst, dessen Prozessierung, seiner Dicke und der Zahl vorhandener Ladungen oder Haftstellen ab. Einen guten Überblick über Leitungsmechanismen in Gatedielektrika von MOSFET-Kondensatoren gibt die Zusammenfassung von Yang *et al.* [YAN04].

Bei der **Schottky-Emission** gelangen Elektronen aus der Gateelektrode durch thermische Anregung über die Potentialbarriere in das Leitungsband des Oxids, wenn das elektrische Feld hoch genug ist. Die angeregten Elektronen tragen dann zum Stromfluss durch das Oxid bei. Das Prinzip der **Poole-Frenkel-Emission** ist dem der Schottky-Emission sehr ähnlich, dabei sind es aber keine Elektronen aus der Gateelektrode, die durch thermische Anregung in das Leitungsband des Oxids gehoben werden, sondern Elektronen, die aus Traps im Gateoxid gelöst werden. Die gelösten Elektronen bilden dann einen Stromfluss durch das Oxid, der stark von der Größe des elektrischen Feldes und der Temperatur abhängt [FRE38]. Schottky- und Poole-Frenkel-Emission finden sich vor allem bei Si_3N_4-Isolatoren, die in Metall-Oxid-Halbleiter-Strukturen Verwendung finden. Siliziumnitrid hat eine kleinere Bandlücke als Siliziumdioxid und beinhaltet häufig eine größere Zahl an Defekten.

Bei thermisch nitridierten Oxiden, die wegen ihrer großen Beständigkeit gegen dielektrische Durchbrüche vor allem dort eingesetzt werden, wo sehr dünne Isolatorschichten notwendig sind, bildet das Tunneln über Haftstellen (engl. „**Trap-assisted tunneling**“) den wesentlichen Leitungsmechanismus. Dabei tunneln die Ladungsträger zunächst zu einer Haftstelle und von dort aus in das Leitungsband des nitridierten Oxids. Aufgrund der Zweiteilung des Leitungsvorgangs wird dieser Mechanismus auch als „Zwei-Schritt-Tunneln“ (engl. „Two-step-tunneling“) bezeichnet [SUZ86], [FLE92].

Weitere Leitungsmechanismen können Phonon-Assisted-Tunneling sowie Leitung durch „heiße“ Ladungsträger (engl. Hot Carrier, siehe auch Kapitel 6.1.4) sein. In Oxiden mit einer Dicke von weniger als 10 nm spielen auch Stress-induzierte Leckströme (SILCs) eine Rolle, die die Oxidqualität nachhaltig beeinflussen können [DUM01].

Der hauptsächliche Ladungsfluss im Siliziumdioxid basiert auf quantenmechanischem Tunneln. Dabei sind zwei Mechanismen zu unterschieden, das **direkte Tunneln** und das **Fowler-Nordheim-Tunneln** [FOW28]. Während Elektronen in dünnen Oxiden (< 5 nm) direkt durch das gesamte Gateoxid tunneln können, ist bei dickeren Oxiden (> 5 nm) eine Absenkung der Potentialbarriere durch eine von außen angelegte Spannung nötig. Dadurch müssen die Elektronen nur noch einen Teil des Oxides durchtunneln, um in das Leitungsband des Oxids zu gelangen. Dieser Mechanismus wird als Fowler-Nordheim-Tunneln bezeichnet. Abbildung 3.2 stellt direktes Tunneln (a) und Fowler-Nordheim-Tunneln (b) gegenüber.

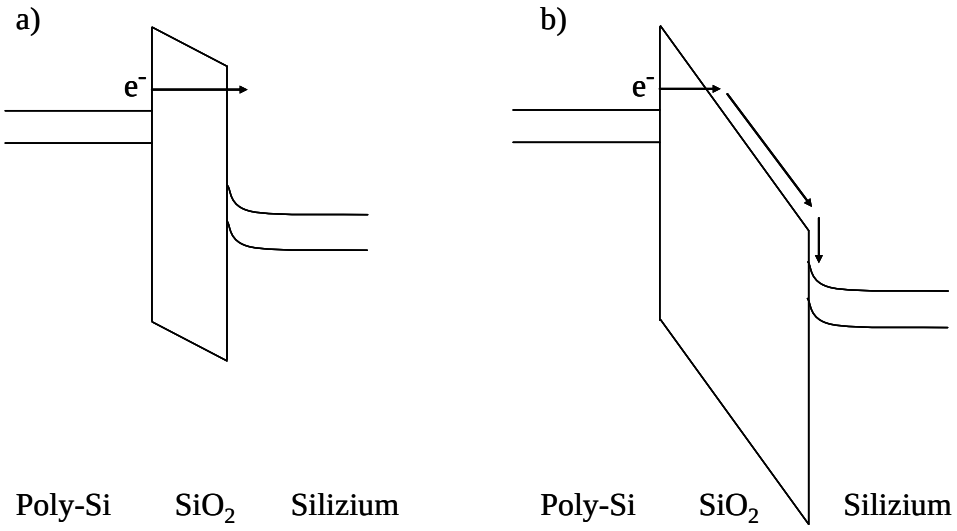


Abbildung 3.2: Gegenüberstellung von direktem Tunneln (a) und Fowler-Nordheim-Tunneln (b) durch das Gateoxid (SiO_2) zwischen Polysilizium-Gate und Siliziumfilm

Fowler-Nordheim-Ströme entstehen beim Anlegen eines elektrischen Feldes. Je höher das elektrische Feld, d. h. je höher die über dem Oxid anliegende Spannung ist, desto größer ist der Fowler-Nordheim-Strom. Die Stromdichte J_{FN} des durch Fowler-Nordheim-Tunneln hervorgerufenen Stromflusses beträgt [PAN95]:

$$J_{FN} = A \cdot E_{ox}^2 \cdot \exp\left(\frac{-B}{E_{ox}}\right) \qquad [3.1]$$

Dabei ist E_{ox} das elektrische Feld und A und B sind Konstanten:

$$A = \frac{e^3 \cdot m_{si}}{8 \cdot \pi \cdot h \cdot m_{ox} \cdot \Phi_b} \quad [3.2]$$

$$B = \frac{4 \cdot \sqrt{2 \cdot m_{ox} \cdot \Phi_b^3}}{3 \cdot (h/2 \cdot \pi) \cdot e} \quad [3.3]$$

In den Formeln 3.2 und 3.3 bezeichnet e die Elementarladung, m_{si} und m_{ox} die effektiven Elektronenmassen im Silizium bzw. im Oxid, h das Plancksche Wirkungsquantum und Φ_b die zu durchtunnelnde Potentialbarriere vom Polysilizium zum Siliziumdioxid. Im JEDEC-Standard JESD92 wird für Silizium und Siliziumdioxid der Quotient m_{ox}/m_{si} mit ungefähr 0,5 angegeben. Weiterhin ist dort $\Phi_b \approx 3{,}1$ eV [JED92]. Für die Konstanten A und B finden sich im JEDEC-Standard JESD35-A typische Werte von 1,6 MA/(MV)² für A und 222 MV/cm für B [JED35].

Mit Gleichung 3.1 ist der klassische Fowler-Nordheim-Strom ohne Einfluss der Temperatur dargestellt. Bei höheren Temperaturen muss die Absenkung des Ferminiveaus und eine dadurch reduzierte Potentialbarriere Φ_b berücksichtigt werden. Die Abhängigkeit des Fowler-Nordheim-Stromes von der Temperatur ist aber nur schwach. Eine analytische Näherung der temperaturabhängigen Fowler-Nordheim-Stromdichte gibt Gleichung 3.4 an [PAN95]:

$$J_{FN}(T) = \frac{\pi \cdot C \cdot k \cdot T}{\sin(\pi \cdot C \cdot k \cdot T)} \cdot J_{FN} \quad [3.4]$$

Hierbei ist k die Boltzmann-Konstante, T die Temperatur in Kelvin, J_{FN} die Fowler-Nordheim-Stromdichte aus Gleichung 3.1 und C folgende Konstante:

$$C = 2 \cdot \frac{\sqrt{2 \cdot m_{ox} \cdot \Phi_b}}{(h/2 \cdot \pi) \cdot e \cdot E_{ox}} \quad [3.5]$$

Der Fowler-Nordheim-Strom steigt mit der Temperatur an. Typische Strom-Spannungs-Kennlinien bei verschiedenen Temperaturen für das 40 nm dicke Gateoxid der H10-Technologie sind in Abbildung 3.3 dargestellt. Bei einem Feld von etwa 6 MV/cm wird der Fowler-Nordheim-Strom erkennbar größer als der Leckstrom. Der Verlauf der Kennlinien ist bei den vier Temperaturen ab etwa 25 V nahezu parallel und der Fowler-Nordheim-Strom bei den höheren Temperaturen etwas größer als bei den tieferen Temperaturen.

Der Fowler-Nordheim-Strom kann Haftstellen erzeugen bzw. diese mit Ladungsträgern füllen. Je mehr solcher Haftstellen entstehen bzw. mit einem Ladungsträger gefüllt sind, desto schlechter werden die Isolatoreigenschaften des Gateoxids. Man spricht in diesem Fall auch von der Degradation des Oxids (engl. oxide degradation oder oxide wearout). Eine fortschreitende Degradation führt irgendwann zum Durchbruch (engl. oxide breakdown, kurz BD). Der Strom kann dann ungehindert durch das Oxid fließen. Abschnitt 3.1.3 führt das Phänomen des Oxiddurchbruchs genauer aus.

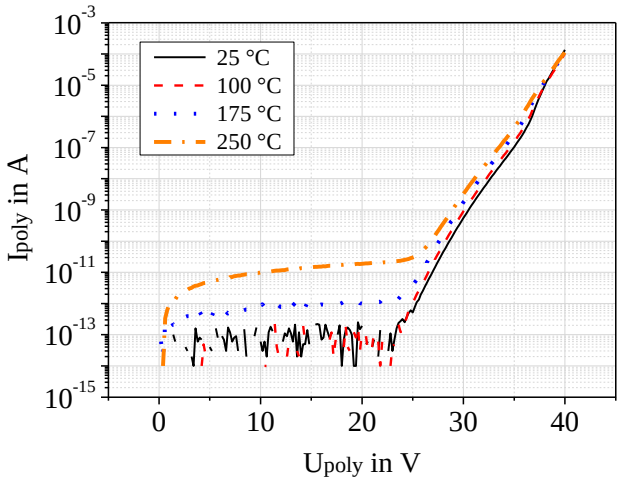


Abbildung 3.3: Strom-Spannungs-Kennlinien des 40 nm dicken Gateoxids bei 25 °C (durchgezogene Linie), 100 °C (gestrichelte Linie), 175 °C (gepunktete Linie) und 250 °C (Strich-Punkt-Linie)

3.1.3 Modelle zum Oxiddurchbruch und zur Lebensdauer von Oxiden

Durch die Zunahme von Haftstellen und den damit verbundenen Stromfluss durch das Oxid kommt es irgendwann zu einem leitenden Pfad durch den Isolator. Man spricht dann von einem dielektrischen Durchbruch. Dabei sind weiche, nicht-zerstörende Durchbrüche (engl. soft breakdown) von harten, zerstörenden Durchbrüchen (engl. hard breakdown) zu unterscheiden. Welcher Durchbruch auftritt, hängt von der Größe der im Kondensator gespeicherten Energie, der Stromdichte am Ort des Durchbruchs sowie der Oxiddicke und der Kondensatorfläche ab. Dielektrische Durchbrüche sind immer lokale Phänomene. Beim weichen Durchbruch entsteht an einer Stelle im Kondensator kurzzeitig eine leitende, aber relativ hochohmige Verbindung zwischen den Elektroden, wodurch der Kondensator an dieser Stelle entladen wird. Die aber weiterhin von außen angelegte Spannung kann den lokalen Kondensator wieder laden, so dass dort der Stromfluss wieder unterbunden wird. Es liegt keine dauerhafte Schädigung des Oxids vor, so dass mehrere weiche Durchbrüche nacheinander auftreten können. Beim harten Durchbruch hingegen ist die Energie an der Stelle, an der eine leitende Verbindung auftritt, so groß, dass eine lokale Erhitzung das Oxid in diesem Bereich zum Schmelzen bringt und dort dann eine dauerhaft leitende Verbindung besteht. Dieser harte Durchbruch ist endgültig und nicht reversibel.

Vor allem bei dickeren Oxiden wie dem des H10-Prozesses ist oft der erste Durchbruch auch gleichzeitig der letzte, da die gespeicherte Energie sehr groß ist. Es konnten aber auch bei 80 nm dicken Oxiden mehrere weiche Durchbrüche hintereinander beobachtet werden [DUM01]. Bei dünnen Oxiden treten häufig viele weiche Durchbrüche vor dem endgültigen harten Durchbruch auf. Bei sehr dünnen Oxiden ist es sogar denkbar, dass nur weiche Durchbrüche stattfinden, da die Energie nie groß genug werden kann, um Bereiche des Oxids zu schmelzen.

Ein einfaches schematisches Modell zum Oxiddurchbruch ist das in Abbildung 3.4 dargestellte **Durchflussmodell** (engl. percolation model). Zu Beginn sind nur wenige, prozessbedingte Haftstellen vorhanden. Durch das Anlegen einer Spannung werden weitere Haftstellen generiert. Die Anzahl der Haftstellen und Ladungen im Oxid nimmt mit der Zeit zu, wodurch sich auch der durch das Oxid fließende Strom verändert. Wenn die Haftstellendichte an einer Stelle groß genug ist, um lokal die Stromdichte stark zu erhöhen, kann ein weicher oder sogar ein harter Durchbruch stattfinden.

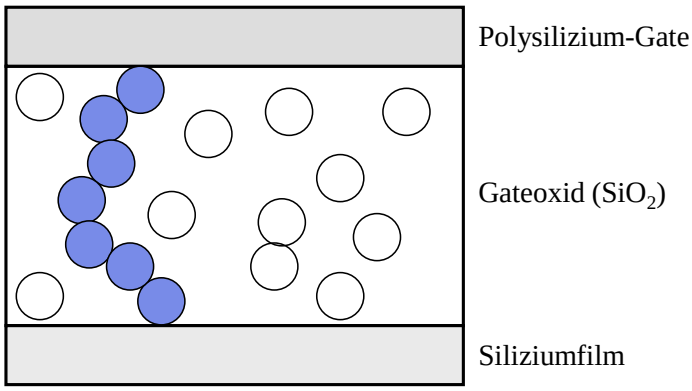


Abbildung 3.4: Durchflussmodell des Oxiddurchbruchs; die den Durchbruch hervorrufenden Haftstellen sind blau eingefärbt, um den Pfad kenntlich zu machen (weitere Haftstellen als leere Kreise)

Für die Zeit bis zum Durchbruch eines bestimmten Oxids ist neben der Größe der angelegten Spannung bzw. des elektrischen Feldes auch die Temperatur ein entscheidender Faktor. Das Durchflussmodell ist aber nur eine qualitative Darstellung und kann deshalb keine genauen Aussagen über den Zusammenhang zwischen Spannung bzw. Feld, Temperatur und zeitlicher Entwicklung bis zum Oxiddurchbruch machen.

Für quantitative Vorhersagen wurden im Laufe der Jahre mehrere Modelle entwickelt. Sie versuchen, die physikalischen Hintergründe des Oxiddurchbruchs zu erklären und zusätzlich Aussagen zur Lebensdauer der Oxide abzuleiten. Dabei haben sich vor allem zwei Modelle hervorgetan, das so genannte E-Modell und das 1/E-Modell, die bei ihrer Vorstellung im Jahr 1985 zu kontroversen Diskussionen geführt haben. Neuere Arbeiten versuchen, beide Modelle zu kombinieren [HU99], [McP01].

Das von McPherson und Baglee [McP85] entwickelte **E-Modell** beruht auf Beobachtungen von Crook [CRO79] und Berman [BER81]. Es geht von einem molekularphysikalischen Ansatz aus und wird deshalb auch als thermochemisches Modell (engl. thermochemical model) bezeichnet. Oxiddegradation entsteht dabei durch die Interaktion des lokalen elektrischen Feldes mit schwachen Si-O-Bindungen der SiO_2-Moleküle. Je nachdem, wie die Moleküle und damit die Si-O-Bindungen zwischen den Molekülen angeordnet sind, ist die Bindung stärker oder schwächer und einfacher zu brechen (siehe auch Abbildung 3.5).

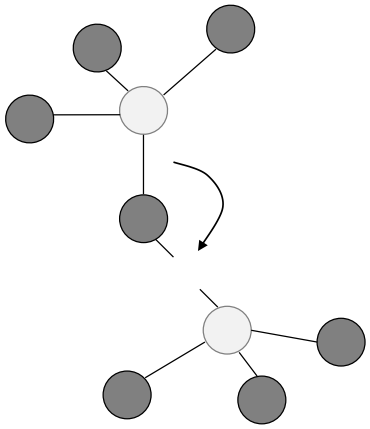


Abbildung 3.5: Lokale Ansicht der Siliziumdioxid-Struktur (SiO_2) mit einer gebrochenen Bindung zwischen einem Siliziumatom (hell) und einem Sauerstoffatom (dunkel)

Werden die offenen Bindungen als Haftstellen interpretiert, so hat eine große Anzahl an offenen Bindungen, hervorgerufen durch ein elektrisches Feld, einen Oxiddurchbruch zur Folge. Für die Zeit bis zum Durchbruch gilt gemäß diesem Modell (Gleichung aus [McP01]):

$$\ln(t_{\mathrm{BD}}) \propto \frac{E_{\mathrm{A(E)}}}{k \cdot T} - \gamma \cdot E_{\mathrm{ox}} \qquad [3.6]$$

Dabei ist t_{BD} die Zeit bis zum Durchbruch, $E_{\mathrm{A(E)}}$, die Aktivierungsenergie, die nötig ist, um eine Bindung zu brechen, k die Boltzmann-Konstante, T die Temperatur, γ der Feldbeschleunigungsfaktor und E_{ox} das elektrische Feld über dem Oxid. Die Proportionalität zum elektrischen Feld gibt dem Modell den Namen.

Das E-Modell gilt für Bindungsenergien kleiner als 3 eV, wenn das Brechen der Oxidbindungen durch feldbeschleunigte thermische Prozesse dominiert wird. Die Vorhersagen des E-Modells stimmen vor allem dann mit experimentellen Ergebnissen überein, wenn das elektrische Feld eher schwach und die Temperaturen hoch sind [McP01].

Im Gegensatz dazu ist das **1/E-Modell**, auch Anoden-Loch-Injektions-Modell (engl. anode hole injection model) genannt, bei größeren Feldern, aber kleineren Temperaturen anzuwenden [CHE85], [LEE88], [MOA89]. Defekte werden hierbei durch so genannte „heiße“ Löcher erzeugt. Über Fowler-Nordheim-Tunneln gelangen Elektronen in das Leitungsband des Oxids. Während sie zur Anode driften, gewinnen sie Energie. Ein Teil dieser Elektronen gibt dann auf Seiten der Anode seine Energie an Valenzbandelektronen ab. Letztere gelangen dadurch in das Leitungsband des Anodenmaterials und lassen im Valenzband „heiße“ Löcher zurück, die wieder in das Oxid zurücktunneln. Dort generieren die Löcher dann Defekte, die als Haftstellen für Elektronen wirken. Der Oxiddurchbruch findet statt, sobald eine kritische Anzahl an Löchern bzw. Haftstellen im Oxid vorhanden ist. Das Prinzip dieses Modells ist schematisch in Abbildung 3.6 dargestellt.

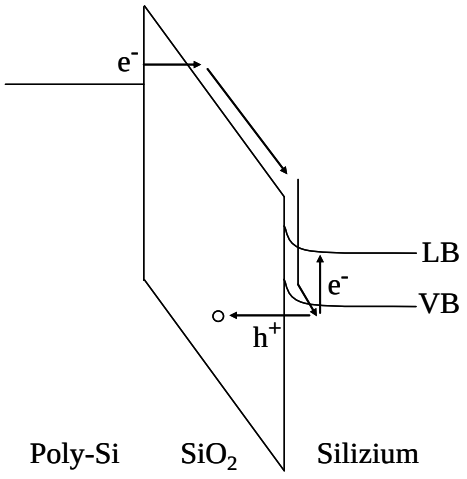


Abbildung 3.6: Schematische Darstellung des 1/E-Modells: Entstehung heißer Löcher, die in das Oxid tunneln und dort für die Erzeugung von Defekten (Haftstellen für Elektronen) verantwortlich sind

Die Lebensdauer eines Oxids, d.h. die Zeit bis zum Durchbruch, wird im 1/E-Modell folgendermaßen beschrieben (Gleichung aus [McP01]):

$$\ln(t_{BD}) \propto \frac{E_{A(1/E)}}{k \cdot T} + G \cdot \frac{1}{E_{ox}} \qquad [3.7]$$

In der Gleichung ist t_{BD} die Zeit bis zum Durchbruch, $E_{A(1/E)}$ die Aktivierungsenergie für die Generierung von Löchern und das Füllen der Haftstellen, k die Boltzmann-Konstante, T die Temperatur, G der Feldbeschleunigungsfaktor und E_{ox} das elektrische Feld über dem Oxid. Die Anti-Proportionalität zum elektrischen Feld bestimmt den Namen des Modells.

Das E-Modell macht pessimistischere Lebensdaueraussagen als das 1/E-Modell und wird daher häufiger angewendet. Auch in den JEDEC-Standards wird das E-Modell zur Ermittlung der Feld-Beschleunigung herangezogen. Bei der Auswertung der Ergebnisse dieser Arbeit wird sowohl die Anwendung des E-Modells als auch die des 1/E-Modells überprüft.

3.1.4 Zeitabhängiger dielektrischer Durchbruch (TDDB)

Um seine Qualität zu bestimmen, wird das Oxid elektrischem Stress in Form von Strom oder Spannung ausgesetzt und Veränderungen im Isolatorverhalten beobachtet. Dabei gibt es zwei wesentliche Szenarien. Der angelegte Strom oder die Spannung kann als Rampe angelegt werden, d. h. Strom bzw. Spannung werden in diskreten Zeitschritten erhöht, bis es zum Durchbruch kommt. Bei Anlegen einer gleichmäßig mit der Zeit laufenden **Spannungsrampe** kann man die Ladung bis zum Durchbruch bestimmen (engl. charge to breakdown, kurz QBD), die ein Indiz für die Qualität des Oxids ist. Solche Tests werden meistens bei Raumtemperatur durchgeführt und dienen nur dem qualitativen Vergleich von Oxiden verschiedener Wafer oder Chargen und nicht der quantitativen Berechnung einer Lebensdauer (siehe auch Kapitel 2.3).

Eine alternative Methode ist das Anlegen eines konstanten Stromes oder einer konstanten Spannung oberhalb der Betriebsbedingungen. In der Praxis wird meistens die **Konstant-Spannungs-Methode** verwendet, da eine konstante Spannung eher einem konstantem elektrischen Feld entspricht als ein konstanter Strom und Gleichung 3.6 oder 3.7 dann anwendbar sind [DUM01]. Liegt bei einer konstanten Temperatur eine konstante Spannung an, wird die Zeit bis zum Oxiddurchbruch gemessen und mit Hilfe des E- oder des 1/E-Modells auf die Lebensdauer des Oxids bei Betriebsbedingungen extrapoliert. Die Konstant-Spannungs-Methode wird als zeitabhängiger dielektrischer Durchbruch (engl. Time-Dependent Dielectric Breakdown, kurz TDDB) bezeichnet. Der Durchbruch ist messtechnisch als plötzlicher Stromanstieg sichtbar.

Die Messmethode TDDB wird im JEDEC-Standard JESD92 [JED92] erläutert. Im JEDEC-Dokument JP001 [JED1] finden sich zudem wertvolle Hinweise zur konkreten Vorgehensweise bei Messung und Auswertung. Bei einer konstanten Stresstemperatur wird an das Oxid eine konstante Stressspannung angelegt und die Zeit bis zum Durchbruch (meistens einem harten Durchbruch) gemessen. Der Durchbruch ist definiert als ein Stromanstieg von einem Faktor 2 bis 10 von einem Messpunkt zum nächsten. Bei dickeren Oxiden ist der harte Durchbruch sichtbar, wenn der Strom in die eingestellte Strombegrenzung weit oberhalb des akzeptierten Strombereichs läuft. Diese Vorgehensweise wird an mehreren identischen Oxiden wiederholt, um eine ausreichende Statistik zu erhalten. Anschließend trägt man die Zeiten bis zum Durchbruch t_{BD} in einem **Weibullgraphen** auf. Abbildung 3.7 zeigt beispielhaft einen solchen Weibullgraphen, dessen Daten aus TDDB-Messungen bei einer konstanten Spannung und einer konstanten Temperatur stammen. Auf der x-Achse sind die Ausfallzeiten t_{BD} und auf der y-Achse die kumulierte Anzahl der Ausfälle W aufgetragen.

W berechnet sich über einen doppelten Logarithmus der kumulierten Ausfallwahrscheinlichkeit *F*:

$$W = \ln\left(-\ln(1-F)\right) \qquad [3.8]$$

Es gibt mehrere, leicht unterschiedliche Methoden, *F* zu berechnen. Eine häufig verwendete Methode ist die von Benard [BEN53], [ABE00]. Dabei bezeichnet *i* die Anzahl der zum Zeitpunkt *t* ausgefallenen Proben und *N* die gesamte Probenzahl zum Zeitpunkt $t = 0$ s.

$$F = \frac{i - 0{,}3}{N + 0{,}4} \qquad [3.9]$$

Aus dem Weibullgraphen kann man bei einem *y*-Wert von Null die Zeit $t_{63\%}$, bis zu der etwa 63 % der Proben ausgefallen sind, ablesen. Alternativ kann man mit Hilfe der Steigung des Weibullgraphen auf die Ausfallzeit von beispielsweise einem ppm ([1]) extrapolieren.

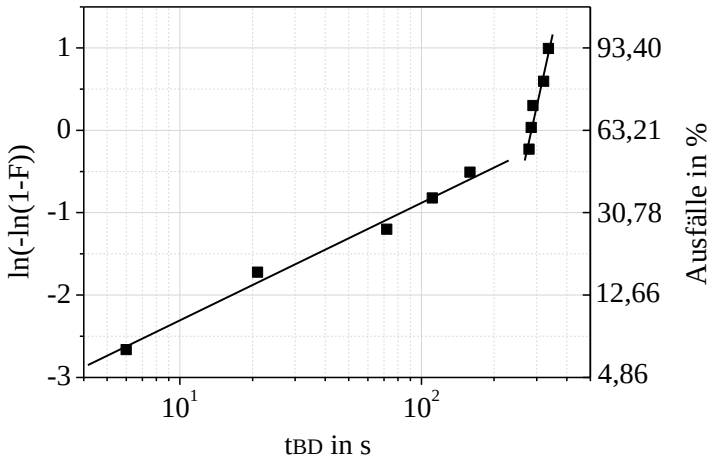


Abbildung 3.7: Beispiel eines Weibullgraphen mit der Anzahl der Oxiddurchbrüche in Abhängigkeit der Ausfallzeiten, die aus TDDB-Messungen gewonnen wurden

Die Weibullverteilung von Oxiddurchbrüchen weist im allgemeinen Fall zwei Bereiche auf. Der Teil mit den kleineren Durchbruchzeiten zeigt eine kleinere Steigung aber eine größere Schwankung. Es handelt sich um den extrinsischen Ast der Verteilung, dessen Lage, Steigung und Schwankungsbreite bei der QBD-Methode von besonderem Interesse sind. Schwache Oxide mit stark prozessbedingten Vorschädigungen wie Defekten, Kontaminationen oder Plasmaschäden haben einen hohen extrinsischen Anteil. Den zweiten Ast mit homogenerer Verteilung, einer größeren Steigung und größeren Ausfallzeiten nennt man den intrinsischen Ast oder Anteil. Er spiegelt ideale Oxide ohne Vorschädigungen wieder, wobei die Statistik der Ausfallzeiten eine Folge des statistischen amorphen Netzwerks des Siliziumdioxids ist. Die Auswertung der TDDB-Messungen zur Bestimmung der Oxid-Lebensdauer bezieht sich auf den intrinsischen Anteil. Ziel der Waferfertigung ist immer, dass der Weibullgraph nur einen intrinsischen Ast hat, d. h. ausschließlich gute Oxide prozessiert werden.

[1] Ein ppm (engl. one part per million) bezeichnet den Anteil von eins zu einer Million.

Werden TDDB-Messungen bei der gleichen Temperatur aber verschiedenen Spannungen, d.h. verschiedenen elektrischen Feldern durchgeführt, so erhält man über das E- bzw. das 1/E-Modell eine Aussage über die Feldbeschleunigung des Oxiddurchbruchs. Dazu trägt man den natürlichen Logarithmus der definierten Zeit bis zum Durchbruch, also z.B. $t_{63\%}$ aus dem Weibullgraphen über dem Feld auf und erhält aus der Steigung den Feldbeschleunigungsfaktor γ bzw. G.

Durch Messungen bei der gleichen Spannung aber unterschiedlichen Temperaturen erhält man die Aktivierungsenergie über das **Arrheniusgesetz** (siehe auch Kapitel 2.3):

$$t_{\mathrm{BD}} = t_0 \cdot \exp\left(\frac{E_{\mathrm{A}}}{k \cdot T}\right) \qquad [3.10]$$

Dabei ist t_{BD} die Zeit bis zum Durchbruch, t_0 eine Zeitkonstante, E_{A} die Aktivierungsenergie, k die Boltzmann-Konstante und T die Temperatur. In einem **Arrheniusgraphen** ist der natürliche Logarithmus der Zeit bis zum Ausfall über $1/(kT)$ aufgetragen. Die Steigung entspricht dann der Aktivierungsenergie.

Der JEDEC-Standard JESD92 empfiehlt Testtemperaturen zwischen 25 °C und 200 °C. Für den H10-Prozess liegt aber schon die typische Betriebstemperatur von 250 °C darüber. Um temperaturbeschleunigte TDDB-Untersuchungen durchführen zu können, die für Zuverlässigkeitsaussagen sehr wichtig sind, müssen also höhere Temperaturen gewählt werden, als es der Standard vorsieht. In dieser Arbeit werden TDDB-Messungen bei Temperaturen zwischen 150 °C und 350 °C durchgeführt und damit der Temperaturbereich zwischen den JEDEC-Vorgaben und den aus H10-Sicht beschleunigenden Temperaturen abgedeckt.

TDDB-Messungen sollten an NMOS- und an PMOS-Kondensatoren[2] jeweils in Inversion und in Akkumulation durchgeführt werden. In Inversion fällt ein Teil der Spannung an der Raumladungszone ab, die effektive Spannung ist also geringer, was tendenziell zu längeren Ausfallzeiten führt. Mit in Akkumulation getesteten Kondensatoren erhält man also kritischere Lebensdaueraussagen, andererseits ist der Inversionsfall der reale Betriebszustand eines Transistors. In den JEDEC-Dokumenten JESD92 und JP001 sind als Mindestanforderung Untersuchungen in Akkumulation vorgeschrieben, aber zur Ergänzung und vollständigen Charakterisierung werden zusätzliche Tests in Inversion empfohlen.

[2] In den JEDEC-Standard werden die Bezeichnungen NMOS- und PMOS-Kondensator verwendet. Auch wenn die Teststrukturen in dieser Arbeit keine Transistoren, sondern tatsächlich nur Kondensatoren ohne Source- und Drain-Gebiet sind, werden auch hier die JEDEC-Bezeichnungen benutzt. Kondensatoren mit n-dotiertem Silizium werden damit zu PMOS-Kondensatoren (wo die n-Dotierung von der Schwellenspannungsimplantation für PMOS-Transistoren realisiert wird) und Kondensatoren mit p-dotiertem Silizium werden damit zu NMOS-Kondensatoren (wo die p-Dotierung von der Schwellenspannungsimplantation für NMOS-Transistoren realisiert wird).

3.2 Durchführung von Tests zur Qualität und Zuverlässigkeit des Gateoxids

3.2.1 Vergleich von NMOS- und PMOS-Kondensatoren in Akkumulation und Inversion

Bevor in Abschnitt 3.2.3 Untersuchungen zur Lebensdauer des H10-Gateoxids vorgenommen werden, sollen zunächst TDDB-Messungen von NMOS- und PMOS-Kondensatoren in Akkumulation und Inversion verglichen werden. Dazu wurden Messungen an Polysilizium-Siliziumdioxid-Silizium-Kondensatoren der Fläche A = 0,2 mm² mit n-dotiertem Polysilizium und n-dotiertem (PMOS) oder p-dotiertem (NMOS) Silizium jeweils mit positiver und mit negativer Gatespannung durchgeführt. Insgesamt ergeben sich also vier Kombinationen: PMOS in Akkumulation (n-Si/U+), PMOS in Inversion (n-Si/U-), NMOS in Akkumulation (p-Si/U-) und NMOS in Inversion (p-Si/U+). Abbildungen der Teststrukturen befinden sich in Anhang A.

In Abbildung 3.8 ist für alle vier Kombinationen der Strom auf Seiten des Polysiliziums in Abhängigkeit von der Zeit für jeweils zehn Messungen dargestellt. Die Stromwerte sind als Absolutwerte aufgetragen, um die logarithmische Darstellung auch für die negativen Ströme zu ermöglichen. Der entsprechende Strom auf der Aktivgebietsseite ist jeweils betragsmäßig gleich groß. Alle Messungen wurden bei 250 °C an Kondensatoren auf demselben Wafer mit einer Gatespannung von ± 38 V durchgeführt. Das elektrische Feld[3] beträgt damit näherungsweise 9,3 MV/cm. Die Gateoxiddicke war jeweils 41 nm. Bei allen Messungen dieses Kapitels lag das Back-Gate-Potential auf 0 V. Der Strom ist in Ampère und nicht als Stromdichte in A/cm² angegeben, um die Daten so zu präsentieren, wie sie gemessen wurden.

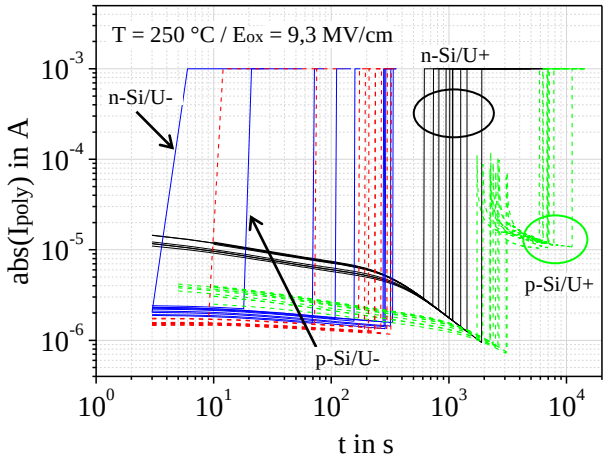


Abbildung 3.8: Strom-Zeit-Kurven bei E_{ox} = 9,3 MV/cm und T = 250 °C in Abhängigkeit von der Polarisation des elektrischen Feldes und der Siliziumdotierung; PMOS in Akkumulation (n-Si/U+), PMOS in Inversion (n-Si/U-), NMOS in Akkumulation (p-Si/U-) und NMOS in Inversion (p-Si/U+); Kondensatorfläche A = 0,2 mm²

Zunächst fällt auf, dass der Strom mit der Zeit abnimmt, bis er einen kritischen Wert erreicht, bei dem ein weicher oder ein harter Oxiddurchbruch stattfindet und der Strom sprunghaft ansteigt. Weiche Durchbrüche konnten nur für die NMOS-Kondensatoren in Inversion (p-

[3] $E_{ox} = U_{ox}/d_{gox}$ mit der Oxiddicke d_{gox} und $U_{ox} = U + U_{FB} + \Phi_S$. Dabei ist U die angelegte Spannung, U_{FB} die Flachbandspannung und Φ_S das Oberflächenpotential. Da Φ_S schwer zu bestimmen ist, kann näherungsweise $E_{ox} = U/d_{gox}$ angenommen werden.

Si/*U*+) festgestellt werden. Nach einer gewissen Zeit erfahren aber auch diese Kondensatoren einen harten Durchbruch. Die Grenze von 1 mA wurde dem Messgerät als Strombegrenzung einprogrammiert. Nachdem einmal die Strombegrenzung erreicht wurde, war das Oxid so weit geschädigt, dass schon bei geringer Spannungsbelastung ein Durchfluss durch das Oxid möglich war.

Wie in Abschnitt 3.1.1 erwähnt, sind Haftstellen in der Nähe der Anode (+) positiv und in der Nähe der Kathode (-) negativ geladen. Die negativ geladenen Haftstellen bzw. der Einfang negativer Ladung ist für die Stromabnahme bis zum Durchbruch verantwortlich [DUM01]. Diese Haftstellen wirken bei hohen angelegten Spannungen, unabhängig von deren Polarisation, als Coulombsche Streuzentren. Zusätzlich schirmen sie das Feld ab, so dass bei konstanter Spannungsbelastung der Strom mit der Zeit sinkt [HAR77].

Weiterhin kann man in Abbildung 3.8 deutlich erkennen, dass sich die gemessenen Durchbruchzeiten stark unterscheiden. Sie erstrecken sich über einen Bereich von weniger als 10 s bis hin zu mehr als 10.000 s, obwohl in allen Fällen Temperatur und angelegte Spannung gleich sind. Betrachtet man den Unterschied zwischen den NMOS- und den PMOS-Kondensatoren, so liegen in Akkumulation die Durchbruchzeiten der PMOS-Kondensatoren (n-Si/*U*+) über den Zeiten der NMOS-Kondensatoren (p-Si/*U*-), in Inversion ist es aber genau umgekehrt (n-Si/*U*- und p-Si/*U*+). Für die PMOS-Kondensatoren sind zudem die Durchbruchzeiten in Akkumulation größer als in Inversion, für die NMOS-Kondensatoren ist es wieder genau entgegengesetzt.

Weibullgraphen

Die Unterschiede werden noch deutlicher, wenn man die zugehörigen Weibullgraphen betrachtet. In Abbildung 3.9 sind mit den Daten aus Abbildung 3.8 die kumulierten Ausfallzahlen *W* in Abhängigkeit von der Durchbruchzeit für die vier Fälle dargestellt. Dabei wurden die kumulierten Ausfallwahrscheinlichkeiten *F* nach der Methode von Benard berechnet (siehe Gleichung 3.9).

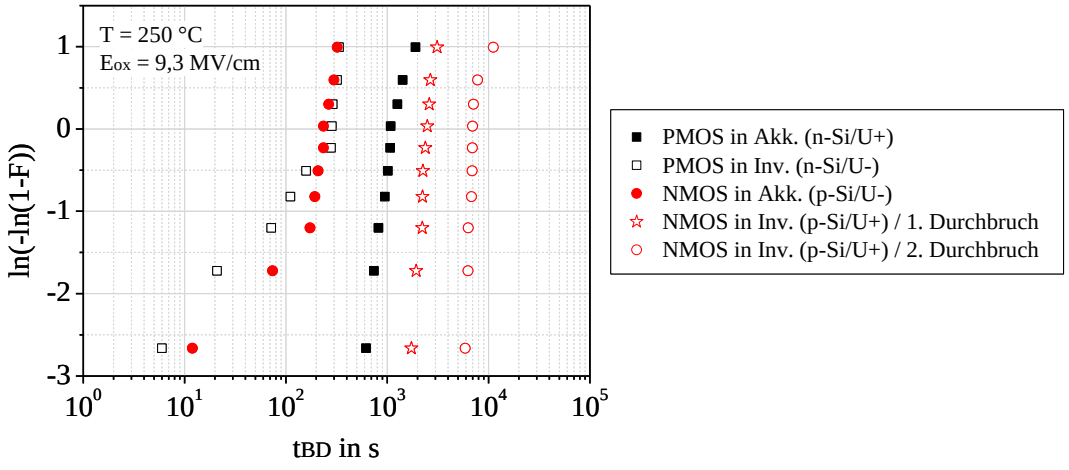


Abbildung 3.9: Kumulierte Anzahl der Ausfälle $W = \ln(-\ln(1-F))$ in Abhängigkeit von der Zeit bis zum Oxiddurchbruch als Weibullgraphen; jede Messreihe besteht aus zehn Messungen bei einer Temperatur von 250 °C und einem konstanten elektrischen Feld von 9,3 MV/cm; Kondensatorfläche A = 0,2 mm²

Auch wenn man erwarten würde, dass aufgrund der zusätzlichen Verarmungsschicht, die den Spannungsabfall über dem Gateoxid reduziert, die Durchbruchzeiten für in Inversion betriebene Kondensatoren immer größer sind als für in Akkumulation betriebene, scheinen hier andere Effekte eine noch größere Rolle zu spielen. In erster Linie sind weder die Dotierung des Siliziums noch Akkumulation oder Inversion alleine ausschlaggebend für die Durchbruchzeiten, sondern die Polarisation des angelegten Feldes. Ist das Feld so gepolt, dass die Polysilizium-Seite die Kathode darstellt (n-Si/*U*- als leere Kästchen und p-Si/*U*- als ausgefüllte rote Kreise), sind bei beiden Silizium-Dotierungen die Durchbruchzeiten wesentlich kürzer als wenn das Polysilizium die Anode ist. Zudem sind extrinsische Äste zu erkennen, was auf prozessbedingte Defekte hinweist. Vergleicht man die Kurven in den Abbildungen 3.8 und 3.9, ist das Stromniveau bei Elektroneninjektion vom Gate, d. h. wenn Polysilizium die Kathode ist, geringer als für Elektroneninjektion vom Substrat.

Zusätzlich muss man aber auch unterscheiden, dass die Durchbruchzeiten bei einem negativen Potential am Polysilizium für NMOS- und PMOS-Kondensatoren ähnlich sind, bei einem positiven Potential am Polysilizium die NMOS-Kondensatoren aber sowohl beim ersten als auch beim zweiten Durchbruch größere Zeiten aufweisen als die PMOS-Kondensatoren. Wenn das Potential am Polysilizium negativ ist, spielt es keine Rolle, ob Inversion oder Akkumulation vorliegt. Wenn aber das Potential am Polysilizium positiv ist, sind für den NMOS-Kondensator die Durchbruchzeiten größer als für den PMOS-Kondensator, weil der NMOS-Kondensator in Inversion ist und der PMOS-Kondensator in Akkumulation. Die zusätzliche Verarmungsschicht reduziert den Spannungsabfall über dem Gateoxid und damit das effektive Feld, so dass der Oxiddurchbruch hinausgezögert wird. Das kleinere effektive elektrische Feld ist in Abbildung 3.8 auch daran zu erkennen, dass das Stromniveau der NMOS-Kondensatoren in Inversion (p-Si/*U*+) kleiner ist als das der PMOS-Kondensatoren in Akkumulation (n-Si/*U*+).

Ladung bis zum Durchbruch

Die Unterschiede im Stromniveau und in den Durchbruchzeiten spiegeln sich auch in der bis zum Durchbruch angesammelten Ladung wider. Integriert man den gemessenen Strom aus Abbildung 3.8 über der Zeit von $t = 0$ s bis $t = t_{BD}$, erhält man die Ladung Q_{BD} bzw. $q_{BD} = Q_{BD} / \mathrm{cm}^2$, die bis zum Durchbruch durch das Oxid fließt. Die Ergebnisse sind in Abbildung 3.10 dargestellt. Normalerweise ist es unüblich, diese Ladung aus TDDB-Messungen zu bestimmen, dafür werden sonst QBD-Messungen durchgeführt (siehe Abschnitt 3.1.4). Hier geht es aber um das Verständnis der Messdaten und nicht um quantitative Vergleiche, deshalb ist die Berechnung der Ladungsmenge sinnvoll.

Während sich in den beiden Fällen der negativen Spannung am Polysilizium bis zum Durchbruch weniger als 0,2 C/cm^2 ansammeln[4], sind dies für den jeweils ersten Durchbruch der beiden anderen Fälle etwa 2 C/cm^2. Betrachtet man den zweiten Durchbruch der NMOS-Kondensatoren in Inversion, werden wegen des hohen Stromniveaus nach dem ersten Durchbruch sogar mehr als 35 C/cm^2 erreicht.

Fasst man die bisherigen Beobachtungen zusammen, ist vor allem die Polarisation der Spannung entscheidend. Ein negatives Potential auf Seiten des Polysiliziums (Elektroneninjektion vom Gate) führt zu einem niedrigeren Stromniveau, weniger Ladung bis zum Durchbruch und kürzeren Durchbruchzeiten als bei einem positiven Potential auf Seiten des

[4] Es ist zu beachten, dass bei negativer Gatespannung ein negativer Strom am Gate gemessen wird und deshalb die berechnete Ladungsmenge auch negativ ist.

Polysiliziums (Elektroneninjektion vom Substrat), zudem werden prozessbedingte Schwachstellen hervorgehoben.

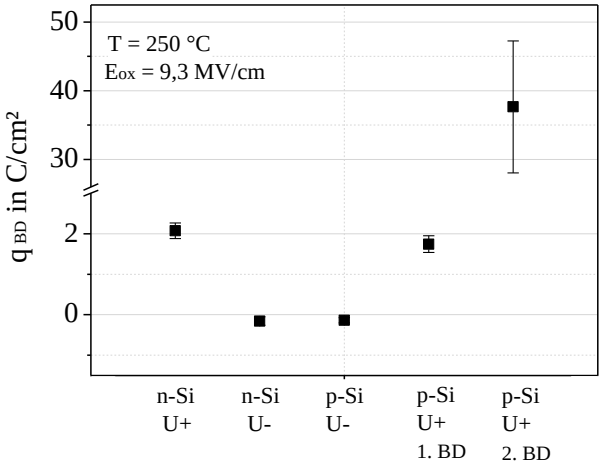


Abbildung 3.10: Ladung pro Fläche beim Oxiddurchbruch in Abhängigkeit von der Polarisation des elektrischen Feldes und der Siliziumdotierung; T = 250 °C, E_{ox} = 9,3 MV/cm; Mittelwerte aus zehn Messungen mit Standardabweichung; bei den negativen Ladungen ist die Standardabweichung so klein, dass die Fehlerbalken nicht sichtbar sind; die Ladung wurde aus dem Integral des gemessenen Stromes über der Zeit bis zum Durchbruch (siehe Abbildung 3.8) bestimmt; Kondensatorfläche A = 0,2 mm²

Die Ursache dafür wird der polykristallinen Struktur des Polysiliziums zugeschrieben, infolge der die Polysilizium-Oxid-Grenzfläche wesentlich rauer ist als die Grenzfläche zwischen dem Oxid und dem einkristallinen Silizium [SCH98]. Eine größere Rauheit führt zu mehr Grenzflächendefekten. Diese Defekte wirken für vom Gate injizierte Elektronen als Haftstellen. Effektiv wird damit die Oxiddicke kleiner oder möglicherweise die Barriere zwischen Polysilizium und Siliziumdioxid reduziert, wodurch sich der Aufbau eines leitenden Pfades durch das Gateoxid vereinfacht. Als Konsequenz erfolgt der dielektrische Durchbruch früher und bei einer kleineren Ladungsmenge. In der Literatur finden sich einige Untersuchungen zu den Auswirkungen rauer Oberflächen, sowohl bei Bulk- als auch bei SOI-Substraten [TUR91], [HON97], [SHE06], mit der Erkenntnis, dass im Falle der Elektroneninjektion von der raueren Elektrode her kleinere Durchbruchzeiten [TSE95] und eine geringere Ladung beim Durchbruch [HON95] auftreten. Entscheidend ist dabei nicht das negative Gatepotential an sich, sondern dass die Elektronen von der Seite der raueren Elektrode her injiziert werden.

Das negative Gatepotential bzw. die Elektroneninjektion vom rauen Polysilizium hat noch weitere Auswirkungen, die in Zusammenhang mit den geringen Durchbruchzeiten und den prozessbedingten Defekten stehen. Darauf wird im nächsten Abschnitt eingegangen.

3.2.2 Strom-Spannungs-Kennlinien der Gateoxid-Kondensatoren

Neben dem zeitabhängigen dielektrischen Durchbruch, aus dem sich die Lebensdauer eines Oxids bestimmen lässt, ist auch die Durchbruchspannung selbst von Interesse. Zum einen gibt sie Auskunft darüber, welche Stressspannung für TDDB-Messungen maximal möglich ist, zum andern kann sie dem schnellen und einfachen Vergleich verschiedener Chargen, Wafer oder auch unterschiedlicher Bereiche auf einem Wafer dienen.

Zur Bestimmung der Durchbruchspannung wird bei konstanter Temperatur eine Spannungsrampe an einen Gateoxid-Kondensator angelegt und der Strom in Abhängigkeit von der Spannung gemessen. Abbildung 3.11 zeigt solche Strom-Spannungs-Kennlinien für PMOS- (a) und NMOS- (b) Kondensatoren für den Fall eines positiven und eines negativen Gatepotentials. Jede Messung wurde einmal wiederholt, so dass die gestrichelten Linien die erste Messung bis zu einer Spannung von ± 35 V und die durchgezogenen Linien die zweite Messung bis zu einer Spannung von mindestens ± 45 V darstellen. Die Tests wurden am selben Wafer durchgeführt, an denen auch die TDDB-Messungen vorgenommen wurden. Die Messungen erfolgten bei 250 °C, verhalten sich qualitativ aber bei Raumtemperatur genauso. Dargestellt sind die Ströme auf der Seite des Polysiliziums, auf der Siliziumseite sind sie betragsmäßig gleich groß.

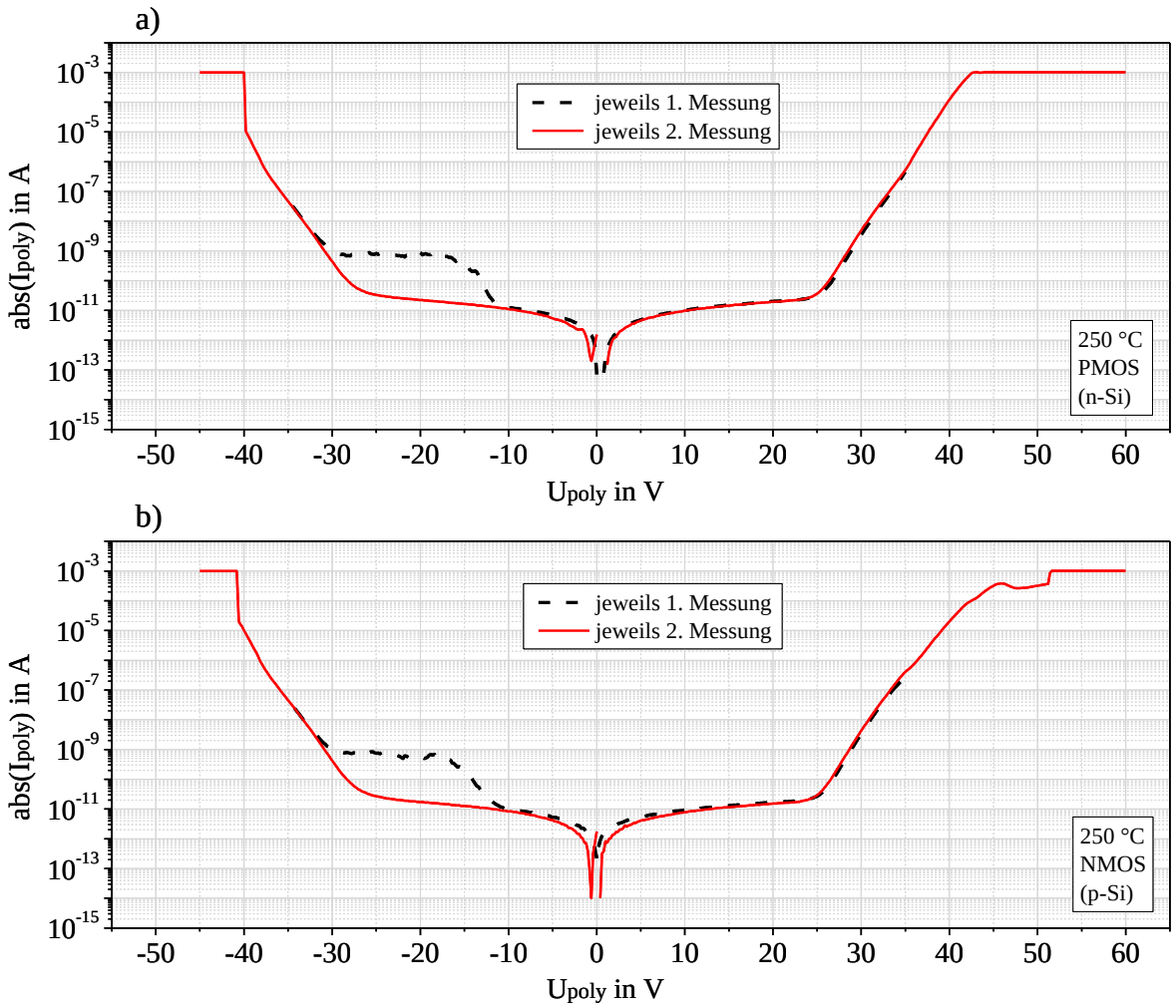


Abbildung 3.11: Strom-Spannungs-Kennlinien von Gateoxid-Kondensatoren mit n-dotiertem (a) bzw. p-dotiertem (b) Silizium; Absolutwerte der Ströme; jeweils zwei Messungen mit positivem und zwei Messungen mit negativem Gatepotential; erste Messung bis ± 35 V (gestrichelte Linien), zweite Messung bis mindestens ± 45 V (durchgezogene Linien); T = 250 °C; Kondensatorfläche A = 0,2 mm²

Es fällt auf, dass bei negativem Gatepotential in der ersten Messung ein „Buckel“ entsteht, d. h. ab einer Spannung von etwa -12 V (entspricht etwa 3 MV/cm) der Strom bis in den Nanoampère-Bereich ansteigt, dann relativ flach weiter läuft und ab etwa -32 V in die Kurve übergeht, die bei der jeweiligen zweiten Messung vorliegt. Erfolgt die erste Messung bis zu einer Spannung von mindestens -32 V, ist weder in der zweiten noch in einer folgenden

Messung ein solcher „Buckel“ sichtbar. Weder das Anlegen eines positiven Gatepotentials für bis zu 24 h noch Erhitzen oder Kühlen des Wafers ruft die „Buckel“ wieder hervor. Bei positivem Gatepotential kann in keinem Fall ein „Buckel“ beobachtet werden.

Bei negativem Gatepotential erfolgt der dielektrische Durchbruch bei beiden Kondensatortypen bei einer Spannung von etwa -40 V. Beim PMOS-Kondensator läuft der Strom bei etwa -42 V in die eingestellte Strombegrenzung von 1 mA, beim NMOS-Kondensator ist dies erst bei etwa -45 V der Fall. Die Spannung, bei der der Fowler-Nordheim-Strom erkennbar größer wird als der Leckstrom, ist in den Fällen des positiven Gatepotentials ungefähr -25 V, was bei einer Gateoxiddicke von 41 nm einem elektrischen Feld von circa 6,1 MV/cm entspricht (siehe auch Abschnitt 3.1.2). In den beiden Fällen des negativen Gatepotentials ist dieses Feld mit circa 6,6 MV/cm leicht höher, trotzdem tritt der Durchbruch früher ein.

Wenn die erste Messung nur bis zu einer Spannung von -20 V erfolgt, ist in der darauffolgenden Messung bis -30 V der „Buckel“ nur bei Spannungen zwischen -20 V und -30 V sichtbar (Abbildung 3.12). In einer dritten Messung bis -40 V ist dann der „Buckel“, der noch in der zweiten Messung bis -30 V zu sehen war, auch nicht mehr vorhanden. Bei dem Auftreten der „Buckel“ handelt es sich nicht um einen Temperatur-Effekt, denn „Buckel“ treten sowohl bei Raumtemperatur als auch bei 250 °C auf. Zudem werden sie bei höherer Temperatur nicht wesentlich größer.

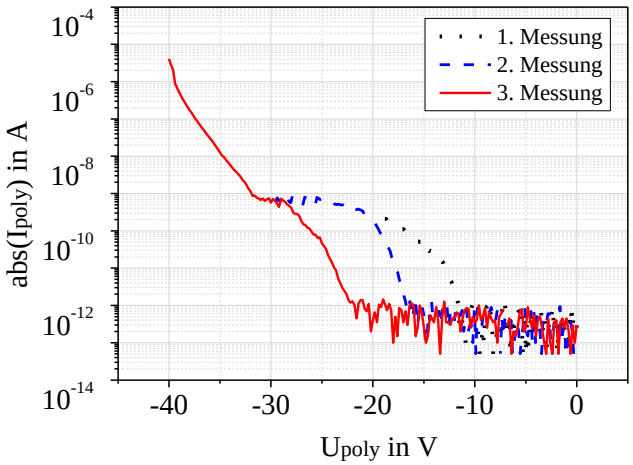


Abbildung 3.12: Strom-Spannungs-Kennlinien eines Gateoxid-Kondensators mit p-dotiertem Silizium; Absolutwerte der Ströme; drei Messungen mit negativem Gatepotential; erste Messung bis -20 V (gepunktete Linie), zweite Messung bis -30 V (gestrichelte Linie); dritte Messung bis -40 V (durchgezogene Linie); Raumtemperatur; Kondensatorfläche A = 0,2 mm²

Die „Buckel“ stehen immer in Zusammenhang mit Elektroneninjektion vom Gate und korrelieren daher mit den Fällen kleiner Durchbruchzeiten und extrinsischer Äste in den Weibullgraphen der TDDB-Messungen. Inversion oder Akkumulation bzw. ob das Silizium n- oder p-dotiert ist, spielt auch hier keine Rolle. Wahrscheinlich sind deshalb auch hier Defekte an der Polysilizium-Siliziumdioxid-Grenzfläche die Ursache. Auch in den bereits erwähnten Untersuchungen in der Literatur traten „Buckel“ bei IU-Kennlinien auf, wenn Elektronen von Seiten der raueren Elektrode her injiziert wurden [TSE95], [HON95], [SHE06].

Die Defekte im Gateoxid an der Polysilizium-Siliziumdioxid-Grenzfläche wirken als Haftstellen. Sie können nur einmal „gefüllt“ bzw. abgesättigt werden (1. Messung, Entstehen der „Buckel“) und geben dann die eingefangene Ladung nicht mehr ab, so dass die „Buckel“ in weiteren Messungen nicht mehr auftauchen. Die eingefangene Ladung wirkt für weitere Ladung abschirmend. Abbildung 3.13 zeigt beispielhaft, wie die Defekte an der Polysilizium-Oxid-Grenzfläche gefüllt werden könnten. Dabei haben die Haftstellen verschiedene Energieniveaus, denn eine negative Gatespannung von -20 V beeinflusst den „Buckel“, der zwischen -20 V und -30 V zu sehen ist, nicht.

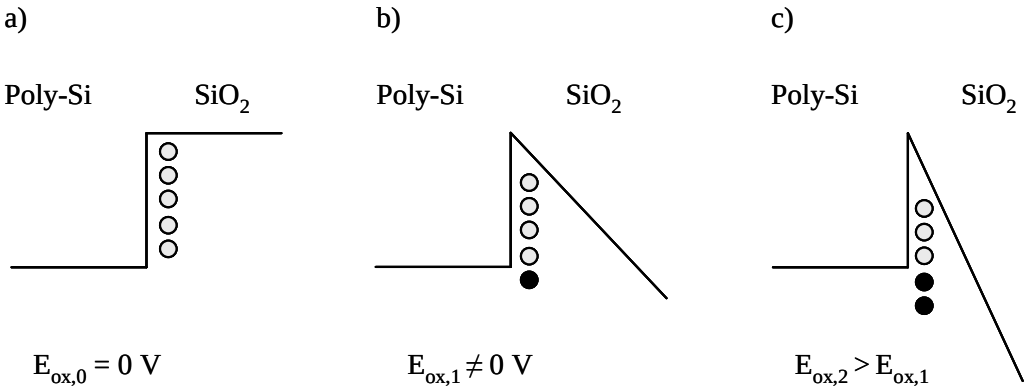


Abbildung 3.13: Schematische Darstellung für das Füllen von Polysilizium-Siliziumdioxid-Grenzflächendefekten; mit Elektronen gefüllte Defekte in schwarz; a) kein Feld liegt an, die Defekte sind unberührt; b) bei Anlegen eines elektrischen Feldes wird zunächst der Defekt im untersten Energieniveau gefüllt; c) wird das Feld erhöht, können auch mehr Defekte gefüllt werden

Ob die Haftstellen sowie die Ladungsträger, die die Defekte füllen, positiv oder negativ geladen sind, ist nicht eindeutig. So könnten direkt Elektronen aus der Polysiliziumelektrode positiv geladene Haftstellen besetzten. Folgt man andererseits der Erklärung von Dumin [DUM01], sind Haftstellen in der Nähe der Kathode aber negativ geladen (siehe auch 3.1.1). Elektronen aus der Polysiliziumelektrode könnten dann das Gateoxid durchqueren und in das Substrat gelangen, wo sie „heiße“ Löcher erzeugen (vergleiche das Prinzip zum 1/E-Modell), die wieder in das Gateoxid zurück tunneln und dort die negativ geladenen Haftstellen besetzen. Da die „Buckel“ nur einmal entstehen und dann nicht wieder auftauchen, muss das Füllen der Defekte oder Haftstellen eine strukturelle Veränderung mit sich ziehen. Beispielsweise könnte eine H-H-Verbindung entstehen, die als Wasserstoffmolekül nun bei der zweiten Messung keinen Einfluss mehr auf den Stromfluss hat. Eine Möglichkeit, dies genauer zu analysieren, wäre eine Untersuchung möglicher Ausgasungen mit Hilfe eines Massenspektrometers.

Als Fazit aus den bisherigen Untersuchungen lässt sich festhalten, dass ein negatives Potential am Polysilizium (PMOS in Inversion und NMOS in Akkumulation) durch zusätzliche Defekte das Ergebnis der TDDB-Messungen negativ beeinflusst. Bei NMOS-Kondensatoren in Inversion findet vor dem harten Durchbruch ein weicher Durchbruch statt und die Durchbruchzeiten und Ladungen bis zum Durchbruch sind höher als in den drei anderen Fällen. Da es in dieser Arbeit nicht um die Beurteilung der Prozessqualität, sondern um die Analyse von Hochtemperatureffekten geht, werden **alle weiteren Untersuchungen** zur quantitativen Analyse der TDDB-Messungen an **PMOS-Kondensatoren in Akkumulation** durchgeführt.

3.2.3 Zeitabhängiger dielektrischer Durchbruch (TDDB) und Abschätzung der Lebensdauer des Gateoxids

TDDB-Messungen wurden an PMOS-Kondensatoren in Akkumulation auf mehreren Wafern der Charge, an denen auch die Messungen zu Abschnitt 3.2.1 vorgenommen wurden, durchgeführt. Die Gateoxiddicke der Kondensatoren betrug jeweils d_{gox} = 41 nm. Die Kondensatoren wurden bei verschiedenen, jeweils konstanten Temperaturen mit unterschiedlichen, jeweils konstanten Spannungen belastet.

In Abbildung 3.14 ist der zeitliche Stromverlauf einiger Messungen dargestellt. Diagramm a) zeigt die Entwicklung des Stromes auf der Seite des Polysiliziums bei 250 °C für Felder zwischen 8,5 MV/cm und 9,6 MV/cm, Diagramm b) den Stromverlauf bei einem Feld von 9,3 MV/cm und Temperaturen zwischen 150 °C und 300 °C. Weitere Messungen wurden bei einem Feld von 8,5 MV/cm und Temperaturen von 200 °C, 300 °C und 350 °C durchgeführt, die sich qualitativ wie die dargestellten Ergebnisse verhalten.

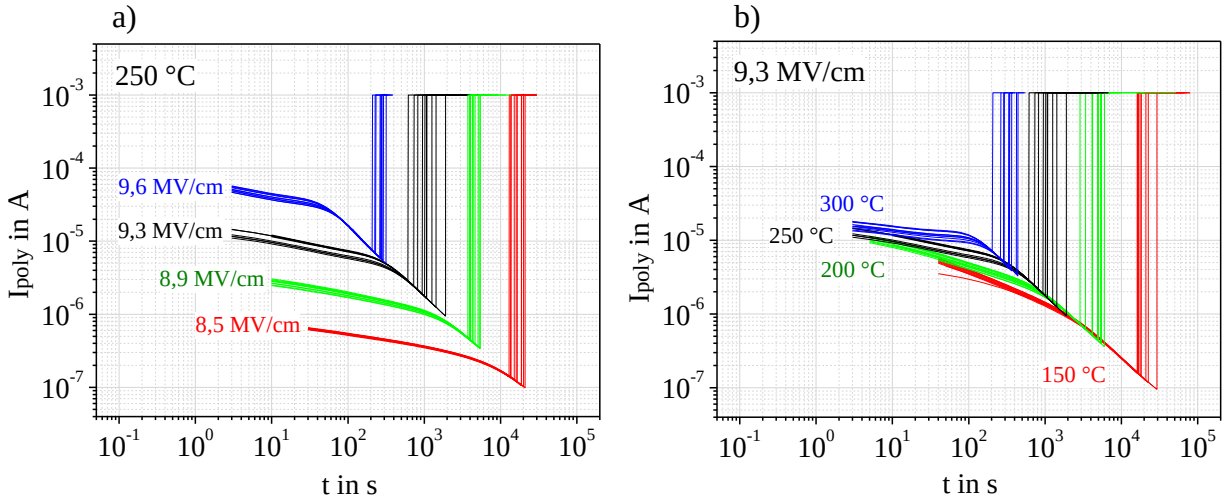


Abbildung 3.14: Strom-Zeit-Kurven der PMOS-Kondensatoren in Akkumulation bei jeweils konstantem elektrischen Feld E_{ox} und konstanter Temperatur T; die Ergebnisse bei 250 °C und Feldern zwischen 8,5 MV/cm und 9,6 MV/cm sind in Diagramm a), die Ergebnisse bei 9,3 MV/cm und Temperaturen zwischen 150 °C und 300 °C in Diagramm b) dargestellt; Kondensatorfläche A = 0,2 mm²

Die Kurven beginnen bei unterschiedlichen Zeiten, weil abhängig von der Dauer bis zum Durchbruch die Zeit zwischen den einzelnen aufgenommenen Messpunkten variiert wurde. Je höher die erwartete Ausfallzeit war, desto größer waren die Zeitabstände zwischen den einzelnen Messpunkten.

Wie schon in Abbildung 3.8 sinkt der Strom wegen der negativ geladenen Haftstellen zunächst mit der Zeit, bevor der Durchbruch stattfindet. Weiche Durchbrüche, d. h. eine zwischenzeitliche leichte Stromerhöhung, konnte in keiner der Messungen festgestellt werden. Das Stromniveau der Kurven ist für größere Felder und für höhere Temperaturen höher, da der Fowler-Nordheim-Strom mit größer werdendem Feld und steigender Temperatur zunimmt. Je größer das Feld und je höher die Temperatur ist, desto schneller erfolgt aber auch der Durchbruch.

Lebensdauer des Gateoxids

Um aus den TDDB-Messungen, die in Abbildung 3.14 dargestellt sind, die Lebensdauer des Gateoxids zu bestimmen, wurden die Ausfallzeiten in Weibullgraphen aufgetragen (Abbildung 3.15).

Extrinsische Äste sind in keinem Fall zu erkennen, d. h. prozessbedingte Schwachstellen liegen nicht vor. Die Fit-Geraden, aus denen die Durchbruchzeiten für die Berechnung der Feld- und Temperaturbeschleunigung gewonnen wurden, sind aus Gründen der Übersichtlichkeit nicht eingezeichnet. Dennoch ist auch so zu erkennen, dass die Steigungen der Kurven im Weibulldiagramm bei den verschiedenen Temperaturen und elektischen Feldern leicht unterschiedlich sind. Oft werden für die weitere Analyse die gemessenen Punkte auf Geraden mit gleicher Steigung für alle Feld-Temperatur-Kombinationen gezwungen. Da dies aber die weiteren Auswertungen beeinflusst und zudem bei der ursprünglichen Beschreibung des E-Modells nicht vorgesehen war [McP85], wird hier darauf verzichtet.

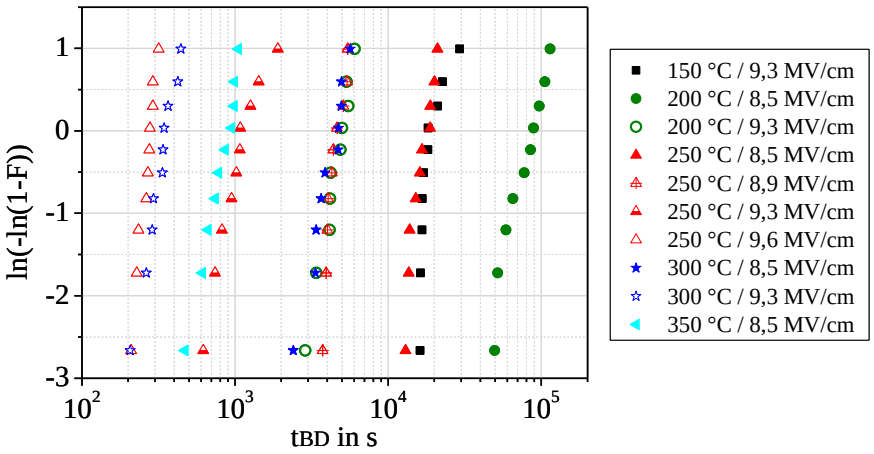


Abbildung 3.15: Kumulierte Anzahl der Ausfälle $W = \ln(-\ln(1-F))$ in Abhängigkeit von der Zeit bis zum Oxiddurchbruch als Weibullgraphen der PMOS-Kondensatoren in Akkumulation; jede Messreihe besteht aus zehn Messungen bei konstanter Temperatur und konstantem elektrischen Feld; Temperaturen von 150 °C bis 350 °C bei Feldern zwischen 8,5 MV/cm und 9,6 MV/cm; die Fit-Geraden sind aus Gründen der Übersichtlichkeit nicht eingezeichnet worden; Kondensatorfläche A = 0,2 mm²

Wie im JEDEC-Standard JESD92 beschrieben, wird die **Feldbeschleunigung** über das E-Modell (siehe Gleichung 3.6) berechnet. Zusätzlich soll die Anwendung des 1/E-Modells (siehe Gleichung 3.7) überprüft werden. Die verwendeten Zeiten $t_{63\%}$ stellen die Ausfallzeiten für den Weibullwert W = 0, d. h. für ca. 63 % der Proben, dar. Abbildung 3.16 zeigt auf der linke Seite (a) den Logarithmus der Ausfallzeiten $t_{63\%}$ in Abhängigkeit von E_{ox} und auf der rechten Seite (b) den Logarithmus der Ausfallzeiten $t_{63\%}$ als Funktion von $1/E_{ox}$.

In dem dargestellten Feld-Bereich sind sowohl das E- als auch das 1/E-Modell auf die Messwerte anwendbar. In beiden Fällen liegen die Steigungen der Kurven bei 200 °C, 250 °C und 300 ° nahe beieinander, ob sie aber temperaturunabhängig sind, lässt sich nicht eindeutig sagen. Beim E-Modell liegen die Messpunkte bei 250 °C ein wenig besser auf einer Geraden als beim 1/E-Modell. Die Datenmenge ist jedoch nicht ausreichend, um dies genau zu klären. Die Feldbeschleunigungsfaktoren γ (E-Modell, Abbildung 3.16 a)) bzw. G (1/E-Modell, Abbildung 3.16 b)) lassen sich aus den Steigungen der Fit-Geraden ablesen. Damit liegt der Faktor γ ungefähr zwischen 3,5 cm/MV und 3,9 cm/MV und der Faktor G zwischen 275 MV/cm und 311 MV/cm.

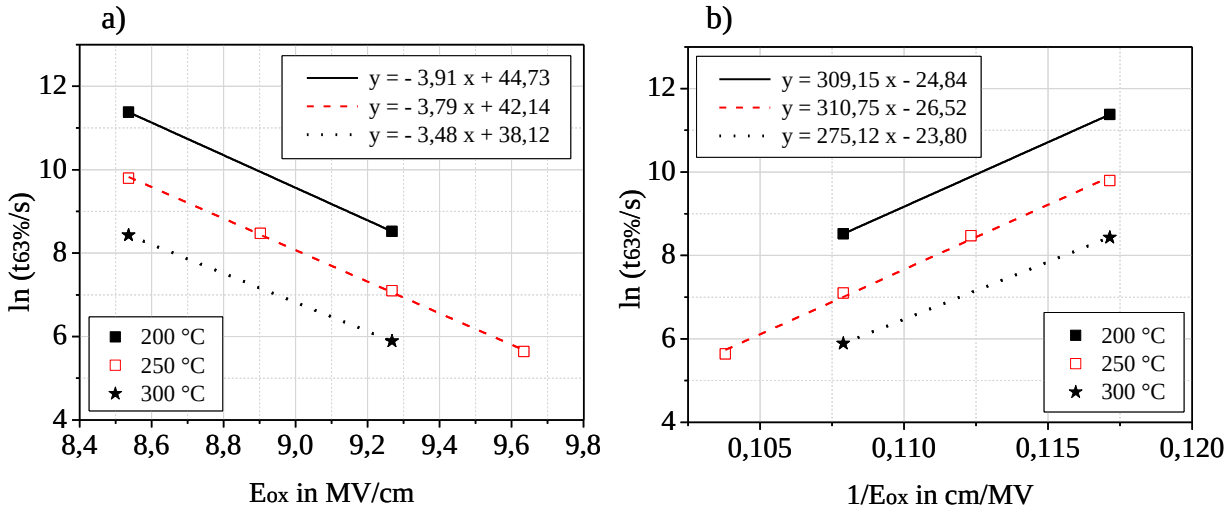


Abbildung 3.16: Logarithmus der Ausfallzeiten bei $W = 0$ (63 % der Proben) in Abhängigkeit vom elektrischen Feld E_{ox} (a) und in Abhängigkeit vom Kehrwert des elektrischen Feldes $1/E_{ox}$ (b); die Messpunkte wurden aus dem Weibulldiagramm in Abbildung 3.15 gewonnen; die Datenpunkte ergeben sich aus jeweils zehn Messungen; $T = 200$ °C (ausgefüllte Kästchen), $T = 250$ °C (leere Kästchen), $T = 300$ °C (Sterne)

In der Literatur finden sich für den Feldbeschleunigungsfaktor γ unterschiedliche Angaben. Die meisten Autoren gehen dabei von einer Temperaturabhängigkeit aus, wie sie auch vom E-Modell vorhergesagt wird. Die von uns ermittelten Werte stimmen mit den Angaben sowohl für den Temperaturbereich bis 150 °C [McP85], [BOY89] als auch für den Bereich oberhalb von 250 °C [YAS99] überein. Für den Faktor G finden sich in der Literatur im Bereich ab 250 °C Werte von etwa 330 MV/cm [MOO07], was ebenso gut mit den hier ermittelten Werten übereinstimmt. Diese Literaturangaben beziehen sich aber alle auf Oxiddicken von höchstens 22 nm. Für eine Oxiddicke von 40 nm konnten Berman *et al.* bei 150 °C einen Feldbeschleunigungsfaktor γ von etwa 3 ermitteln [BER81].

Die **Temperaturbeschleunigung** wurde mit Hilfe des Arrheniusgesetztes (siehe Gleichung 3.10) ermittelt. In Abbildung 3.17 ist der Logarihtmus der Ausfallzeiten $t_{63\%}$ in Abhängigkeit von $1/(kT)$ dargestellt. Für beide elektrischen Felder ergibt sich im Temperaturbereich von 150 °C bis 300 °C (E_{ox} = 9,3 MV/cm) bzw. von 200 °C bis 350 °C (E_{ox} = 8,5 MV/cm) ein linearer Zusammenhang zwischen $\ln(t_{63\%})$ und $1/(kT)$.

Aus den Steigungen dieser Fit-Geraden lassen sich die Aktivierungsenergien ablesen. Die beiden Werte von etwa 0,57 eV bei E_{ox} = 9,3 MV/cm und etwa 0,76 eV bei E_{ox} = 8,5 MV/cm liegen in einem Bereich, der auch bei Untersuchungen in der Literatur für Felder zwischen 8 MV/cm und 10 MV/cm ermittelt werden konnte, sowohl im Temperaturbereich bis 150 °C [VIN96] als auch oberhalb von 250 °C [SUE94], [SUE97], [MOO07]. Dort sinkt die Aktivierungsenergie mit höher werdendem elektrischem Feld, wie es auch bei uns der Fall ist und vom E-Modell vorhergesagt wird [McP85].

Für das hier vorliegende Oxid ist es schwierig, den Bereich des elektrischen Feldes von 8,5 MV/cm bis 9,6 MV/cm weiter auszudehnen. Bei der Messung von IU-Kennlinien liegt bei einer Temperatur von 250 °C der Strom bei einer Spannung von etwa 45 V schon bei 1 mA, so dass elektrische Felder von mehr als 11 MV/cm gar nicht möglich sind. Größere elektrische Felder können vor allem bei dünneren Oxiden verwendet werden. Auf der anderen Seite bedeuten kleinere Felder, dass die Messzeiten länger werden. Für das Feld von 8,5 MV/cm

liegen die Durchbruchzeiten bei 200 °C schon im Bereich bis 100.000 s. Viel kleinere Felder und damit noch größere Messzeiten sind schwer realisierbar.

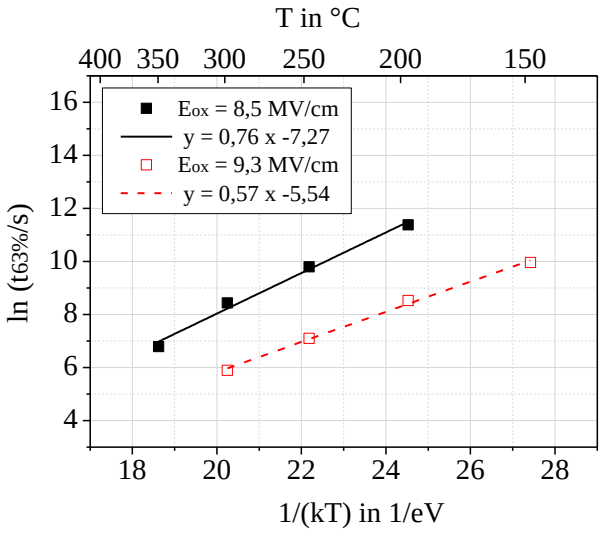


Abbildung 3.17: Logarithmus der Ausfallzeiten bei $W = 0$ (63 % der Proben) in Abhängigkeit von $1/(kT)$; die Messpunkte wurden aus dem Weibulldiagramm in Abbildung 3.15 gewonnen; $E_{ox} = 8{,}5$ MV/cm (ausgefüllte Kästchen), $E_{ox} = 9{,}3$ MV/cm (leere Symbole); die Datenpunkte ergeben sich aus jeweils zehn Messungen

Die mögliche Temperaturabhängigkeit der Feldbeschleunigungsfaktoren und die Feldabhängigkeit der Aktivierungsenergie erschweren es im Allgemeinen, korrekt auf die Betriebsbedingungen zu extrapolieren und damit eine Lebensdauer für das Gateoxid zu erhalten. Da hier aber der Feldbeschleunigungsfaktor auch bei der maximalen Betriebstemperatur von 250 °C ermittelt wurde, fällt die Notwendigkeit einer Temperaturextrapolation weg und die mögliche Temperaturabhängigkeit des Feldbeschleunigungsfaktors spielt keine Rolle, so dass eine Abschätzung der Lebensdauer des H10-Gateoxids bei 250 °C vorgenommen werden kann.

Um nun diese Lebensdauer zu bestimmen, kann mit Hilfe von Gleichung 3.6 (E-Modell)[5] und den Ergebnissen aus Abbildung 3.16 auf die maximal erlaubte Betriebsspannung von 12 V extrapoliert werden. Bei einer Gateoxiddicke von 40 nm entspricht dies einem Feld von 3 MV/cm. Dabei muss berücksichtigt werden, dass sich die ermittelten Beschleunigungsfaktoren auf eine Gateoxidfläche von 0,2 mm^2 beziehen. Für eine typische Schaltung der H10-Technologie wird aber eine Gateoxidfläche von etwa 85 mm^2 benötigt (ca. 8 % der gesamten Schaltungsfläche). Kennt man die Ausfallzeit $t_{BD,1}$ bei Betriebsbedingungen für $A_1 = 0{,}2$ mm^2, kann man mit Hilfe des folgenden, im JEDEC-Standard JESD92 angegebenen Zusammenhangs auf die TDDB-bedingte Ausfallzeit $t_{BD,2}$ der Schaltung mit $A_2 = 85$ mm^2 extrapolieren:

$$\frac{t_{BD,1}}{t_{BD,2}} = \left(\frac{A_2}{A_1}\right)^{1/\beta} \qquad [3.11]$$

Dabei ist β die Steigung einer Geraden im Weibullgraphen, wenn auf der x-Achse $\ln(t_{BD})$ aufgetragen ist. β liegt zwischen 3,7 (200 °C / 8,5 MV/cm) und 8,6 (250 °C / 9,6 MV/cm).

[5] Da das 1/E-Modell weniger kritische Lebensdaueraussagen macht, wird nur mit dem E-Modell extrapoliert.

Die Zeit, nach der ein ppm einer Gateoxidfläche von 0,2 mm^2 bei 250 °C und 3 MV/cm ausgefallen sind, beträgt dann für Feldbeschleunigungsfaktoren γ zwischen 3,48 und 3,91 zwischen 11.000 und 160.000 Jahren. Als untere Abschätzung der Lebensdauer für eine typische H10-Schaltung erhält man dann immer noch mehr als 2000 Jahre. Diese Werte scheinen unrealistisch, aber legt man die von einigen Autoren ermittelten Temperatur- und Feldbeschleunigungsfaktoren zu Grunde, erhält man auch solche Lebensdauerzeiten [MOO07], [VIN96], [YAS99].

Insgesamt fügen sich die Ergebnisse des H10-Gateoxids für die Feldbeschleunigungsfaktoren, Aktivierungsenergien und die Lebensdauer bei Betriebsbedingungen gut in bisherige, in der Literatur beschriebene Untersuchungen ein. Es wurde dabei festgestellt, dass in dem untersuchten Feldbereich das E-Modell und das 1/E-Modell sowie das Arrheniusgesetz auch bei Temperaturen oberhalb der im JEDEC-Standard vorgesehenen 200 °C anwendbar sind. Es ist also möglich, für Technologien mit hohen Betriebstemperaturen beschleunigte TDDB-Messungen bis mindestens 350 °C durchzuführen.

3.2.4 Analyse der Durchbruchbedingungen

In Abbildung 3.14 sind neben der Lebensdauer des Gateoxids bei Betriebsbedingungen noch weitere Informationen bezüglich des Durchbruchverhaltens enthalten. Wie schon erwähnt, ist der Stromwert beim ersten abgebildeten Messpunkt bei einem größeren elektrischen Feld oder einer höheren Temperatur größer und die Zeit bis zum Durchbruch geringer als bei einem kleineren Feld bzw. einer niedrigeren Temperatur. Bei höheren Temperaturen sinkt zudem der Strom bis zum Durchbruch weniger stark als es bei niedrigeren Temperaturen der Fall ist. Der Strom, bei dem das Oxid durchbricht, hängt also stark von der Temperatur und dem elektrischen Feld ab. Abbildung 3.18 verdeutlicht den Zusammenhang zwischen dem Durchbruchstrom I_{BD} und der Durchbruchzeit t_{BD} zusammen für alle Daten der verschiedenen Temperatur-Feld-Kombinationen. Dabei ist I_{BD} definiert als der letzte Strommesswert vor dem Erreichen der Strombegrenzung, d. h. vor dem Durchbruch.

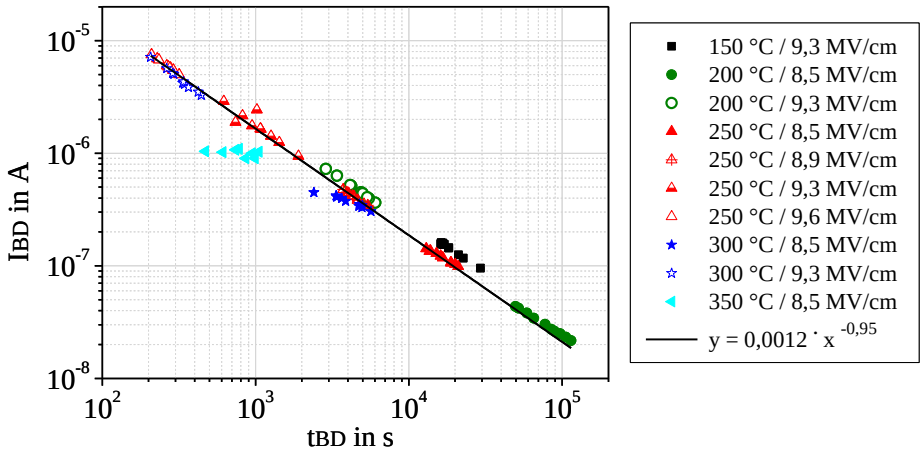


Abbildung 3.18: Durchbruchstrom I_{BD} beim Oxiddurchbruch in Abhängigkeit von der Zeit t_{BD} bis zum Durchbruch bei elektrischen Feldern E_{ox} zwischen 8,5 MV/cm und 9,6 MV/cm und Temperaturen von 150 °C bis 350 °C; die Symbole und Farben wurden analog zu Abbildung 3.15 gewählt; die Fit-Gerade bezieht sich auf alle Daten außer den Werten bei 350 °C; Messdaten aus Abbildung 3.14

Für den Zusammenhang zwischen dem Durchbruchstrom und der Durchbruchzeit gilt ein Potenzgesetz, unabhängig vom elektrischen Feld und der Temperatur:

$$I_{BD} = A \cdot (t_{BD})^{-B} \qquad [3.12]$$

A und *B* sind Konstanten, die mindestens im Temperaturbereich von 150 °C bis 300 °C und im Feldbereich von 8,5 MV/cm bis 9,6 MV/cm temperatur- und feldunabhängig sind. Aus der Fit-Geraden in Abbildung 3.18 kann man für die Konstante *B* einen Wert von etwa 1 ablesen, d. h. es gilt annähernd:

$$I_{BD} \propto \frac{1}{t_{BD}} \quad \text{bzw.} \quad I_{BD} \cdot t_{BD} = \text{konstant} \qquad [3.13]$$

Dieser Zusammenhang findet sich auch in Diagrammen aus der Literatur [DUM01], [PEN94] wobei dabei keine Erklärung für den Zusammenhang gegeben wird. Gleichung 3.13 bedeutet, dass das Produkt aus dem Durchbruchstrom und der Durchbruchzeit konstant ist. Es ist bemerkenswert, dass die Durchbruchströme aus Abbildung 3.14 nicht nur in dieselbe Steigung laufen, sondern sogar in eine gemeinsame Gerade.

In Abbildung 3.19 ist als Ergänzung zu Abbildung 3.18 die Abhängigkeit des Durchbruchstromes I_{BD} vom elektrischen Feld (a) und vom Kehrwert der Temperatur (b) dargestellt. Mit Gleichung 3.6 und Gleichung 3.13 sowie Gleichung 3.10 und Gleichung 3.13 ergeben sich exponentielle Zusammenhänge zwischen dem Durchbruchstrom und dem elektrischen Feld bzw. dem Durchbruchstrom und dem Kehrwert der Temperatur, was durch Abbildung 3.19 bestätigt werden kann. Der Durchbruchstrom I_{BD} nimmt dabei im betrachteten Feldbereich bei 250 °C exponentiell mit dem elektrischen Feld zu (a) und sinkt exponentiell mit dem Kehrwert der Temperatur ab (b).

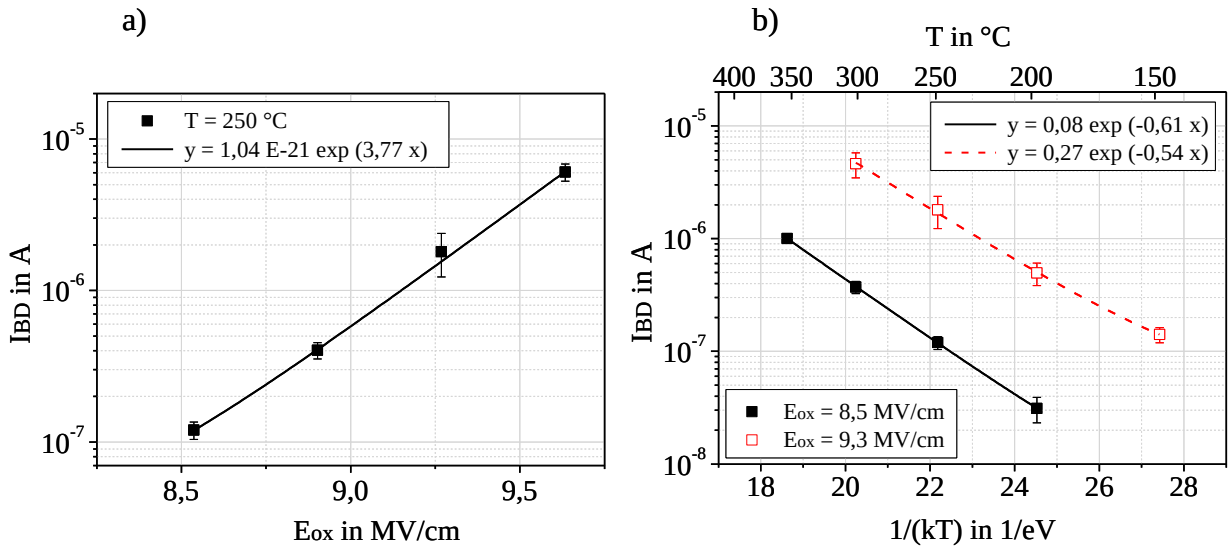


Abbildung 3.19: Durchbruchstrom I_{BD} beim Oxiddurchbruch in Abhängigkeit vom elektrischen Feld E_{ox} bei konstanter Temperatur T = 250 °C (a) und in Abhängigkeit von der Temperatur bei konstantem elektrischem Feld (b) von 8,5 MV/cm (ausgefüllte Kästchen) und 9,3 MV/cm (leere Kästchen); Mittelwerte aus zehn Messungen mit Standardabweichung; Messdaten aus Abbildung 3.14

Neben der Abhängigkeit des Durchbruchstromes von der Zeit bis zum Durchbruch (Abbildung 3.18) sind auch für die Zusammenhänge zwischen dem Durchbruchstrom und dem elektrischen Feld sowie dem Durchbruchstrom und der Temperatur (Abbildung 3.19), die aus TDDB-Messungen gewonnen wurden, in der Literatur keine Angaben zu finden. Es muss aber auch darauf hingewiesen werden, dass bei TDDB-Messungen normalerweise vor allem die Lebensdauer für bestimmte Temperatur-Feld-Bedingungen und nicht der Durchbruchstrom von Interesse ist.

Obwohl mit Gleichung 3.13 ein Zusammenhang zwischen dem Durchbruchstrom und der Durchbruchzeit besteht, der unabhängig vom elektrischen Feld und der Temperatur ist, muss aber damit in der Regel nicht auch die Ladung Q_{BD} bis zum Durchbruch konstant sein, da für die Berechnung der Ladungsmenge als zeitliches Integral über dem Strom auch der Verlauf dieses Stromes bis zum Durchbruch relevant ist. Abbildung 3.20 zeigt dazu die bis zum Durchbruch geflossene Ladung $q_{BD} = Q_{BD} / \mathrm{cm}^2$ als Funktion des elektrischen Feldes bei 250 °C (a) und als Funktion der Temperatur bei Feldern von 8,5 MV/cm bzw. 9,3 MV/cm (b).

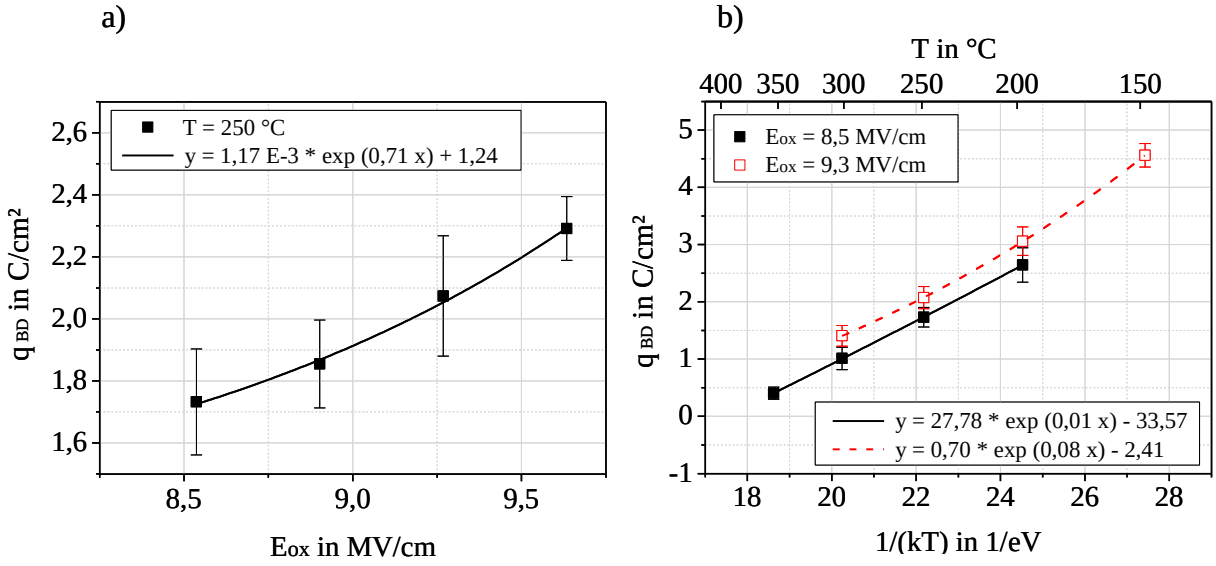


Abbildung 3.20: Ladung pro Fläche beim Oxiddurchbruch in Abhängigkeit vom elektrischen Feld E_{ox} bei konstanter Temperatur T = 250 °C (a) und in Abhängigkeit von der Temperatur bei konstantem elektrischem Feld (b) von 8,5 MV/cm (schwarze Kästchen) und 9,3 MV/cm (leere Kästchen); Mittelwerte aus zehn Messungen mit Standardabweichung; die Ladung wurde aus dem Integral des gemessenen Stromes über der Zeit bis zum Durchbruch bestimmt; Messdaten aus Abbildung 3.14

In Abbildung 3.20 a) nimmt die detektierte Ladungsmenge im untersuchten Bereich exponentiell mit dem elektrischen Feld zu. Entweder ist also bei einem höheren Feld auch eine höhere Ladung nötig, um den Oxiddurchbruch hervorzurufen, oder nur ein Teil der berechneten Ladungsmenge trägt tatsächlich zum Durchbruch bei und ein anderer Teil sammelt sich zwar an, führt aber zu keiner Schädigung. Ein solcher Zusammenhang, wie er in Abbildung 3.20 a) dargestellt ist, lässt sich in der Literatur für ein 40 nm dickes Oxid bei 250 °C nicht finden. Bei dünneren Oxiden (ca. 5 bis 10 nm) findet sich hingegen der umgedrehte Zusammenhang, d. h. dort nimmt die Ladung mit steigendem Feld exponentiell ab [SCH94], [SCH94b], [YEO01], [STA01].

Das Temperaturverhalten der Ladung beim Durchbruch ist genau umgekehrt zum Zusammenhang zwischen Ladung und Feld. Mit steigender Temperatur sinkt die Ladungsmenge, die bis zum dielektrischen Durchbruch durch das Oxid fließt. Bei einer höheren Temperatur ist der

Stromfluss größer, so dass Ladungen einfacher durch das Oxid gelangen und auch einfacher Haftstellen nutzen können, so dass schon eine geringere Anzahl an Defekten für den harten Oxiddurchbruch ausreicht. Der Zusammenhang zwischen der Ladung und dem Kehrwert der Temperatur ist in einem Bereich von 150 °C bis 350 °C exponentiell. In der Literatur wurde im Temperaturbereich bis 200 °C an Oxiden bis 5 nm Dicke ebenso eine Abnahme der Ladung mit der Temperatur beobachtet und auch ein exponentieller Zusammenhang festgestellt [DiM99], [WU01].

Insgesamt lässt sich festhalten, dass die Ladung, die bis zum Durchbruch durch das Oxid fließt, nicht allein für den dielektrischen Durchbruch eines Isolators ausschlaggebend ist, da sie vom elektrischen Feld und der Temperatur abhängt. Es gibt aber mit der Tatsache, dass das Produkt aus dem Strom beim Durchbruch und der Zeit bis zum Durchbruch konstant ist (Gleichung 3.13), einen allgemeinen Zusammenhang, dem alle Durchbrüche folgen, unabhängig davon, welches Feld und welche Temperatur den Durchbruch hervorgerufen haben.

Fazit

Bei den Untersuchungen des 40 nm dicken Gateoxids der H10-Technologie konnten im Temperaturbereich bis 350 °C einige wichtige Beobachtungen gemacht und Schlüsse gezogen werden, die Einfluss auf die Zuverlässigkeit des Gateoxids und dessen Analyse haben. Sie werden im Folgenden zusammengefasst:

- TDDB-Messungen hängen stark von der Polarisation des elektrischen Feldes ab. Kleine Durchbruchzeiten, extrinsische Ausfälle und eine geringe Ladungsmenge bis zum Durchbruch liegen dann vor, wenn bei der Messung Elektronen vom Gate injiziert werden. Dabei ist nicht die Gateinjektion an sich problematisch, sondern die Injektion vom rauen Polysilizium her. Es handelt sich dabei nicht um einen SOI-spezifischen Effekt.

- Raue Oberflächen erzeugen Defekte an der Grenzfläche zum Oxid, die die Gateoxidqualität beeinflussen. Dies wird auch in IU-Kennlinien deutlich, wo „Buckel“ bei der Messung eines zuvor ungestressten Kondensators auftreten. Diese „Buckel“ sind von der Größe der angelegten Spannung abhängig, treten bei weiteren Messungen aber nicht mehr auf. Die Defekte an der Polysilizium-Oxid-Grenzfläche können also nur einmal abgesättigt werden und schirmen dann weitere Ladung ab.

- Für TDDB-Messungen an PMOS-Kondensatoren in Akkumulation stimmen die ermittelten Feldbeschleunigungsfaktoren und Aktivierungsenergien mit Angaben in der Literatur überein. Bei den TDDB-Messungen war es zudem möglich, den im JEDEC-Standard vorgesehenen Temperaturbereich von 200 °C auf 350 °C auszudehnen. Die mit Hilfe des E-Modells ermittelte Lebensdauer von einem ppm der Gateoxidfläche einer typischen H10-Schaltung ist, wie in der Literatur häufig zu finden, mit mehr als 2000 Jahren sehr hoch.

- Es gibt einen universellen Zusammenhang zwischen dem Durchbruchstrom und der Zeit bis zum Durchbruch ($I_{BD} \cdot t_{BD} \approx$ konstant), der für das untersuchte Gateoxid mindestens im Temperaturbereich von 150 °C bis 300 °C für Felder zwischen 8,5 MV/cm und 9,6 MV/cm gültig ist. Die Ladung, die bis zum Durchbruch durch das Oxid fließt, ist nicht das alleinige Durchbruchskriterium, da sie vom elektrischen Feld und der Temperatur abhängt.

Kapitel 4

Zuverlässigkeitsuntersuchungen an EEPROM-Speicherzellen

Nicht-flüchtige Speicher wie EEPROMs sind eine Schlüsseltechnologie für flexible Datenspeicherung, beispielsweise von Kalibrier- und Messinformationen. Zwei Zuverlässigkeitsaspekte sind dabei von entscheidender Bedeutung: der Datenerhalt über einen möglichst langen Zeitraum sowie die so genannte Datenwechselstabilität oder Zyklenfestigkeit. Beide Aspekte sind wegen der zunehmenden Oxiddegradation unter hohen Einsatztemperaturen stark eingeschränkt.

Im vorliegenden Kapitel geht es um die Leistungsfähigkeit der EEPROM-Speicher auf SOI-Substraten bei Temperaturen bis 450 °C und um die Frage, wie quantitative Vorhersagen über die Zuverlässigkeit und Lebensdauer der Speicher gemacht werden können. Für die Analysen wurde hauptsächlich das Verhalten von einzelnen EEPROM-Zellen untersucht. Zusätzliche Tests an EEPROM-Arrays dienen einer statistischen Untermauerung der Ergebnisse.

4.1 Theoretische Aspekte zu EEPROM-Speicherzellen

4.1.1 Halbleiterspeicher

Wie an alle Speicher werden auch an Halbleiterspeicher zahlreiche Anforderungen wie schnelles Schreiben und Lesen, lange Speicherzeiten, geringer Spannungs- und Platzbedarf, große Speicherkapazitäten und geringe Herstellungskosten gestellt. Kein Speicher kann alle Anforderungen gleichermaßen erfüllen, aber nicht immer sind die verschiedenen Aspekte gleich relevant. Deshalb gibt es eine Vielzahl unterschiedlicher Halbleiterspeicher, wobei für jeden Bedarf ein anderer Typ sinnvoll ist.

Halbleiterspeicher werden in zwei Gruppen aufgeteilt: „flüchtige“ und „nicht-flüchtige“ Speicher. Zu den „flüchtigen“ Halbleiterspeichern, die die auf ihnen gespeicherte Information bei Abschalten der äußeren Spannung verlieren, gehören das Static Random-Access Memory (SRAM) und das Dynamic Random-Access Memory (DRAM). Das SRAM ermöglicht schnelles Schreiben und Lesen, hat aber einen gewissen Platzbedarf. Das DRAM benötigt eine kleinere Fläche, ist aber technologisch aufwendiger herzustellen. Zur Gruppe der „nicht-flüchtigen“ Speicher gehören vor allem die Read Only Memory (ROM) - Speicher. Diese Bauelemente verlieren ihre Ladung bei definierten Umgebungsbedingungen auch beim Abschalten der Versorgungsspannung nicht. Das ROM und das programmierbare ROM (PROM) sind schnell im Lesen und benötigen wenig Fläche, können aber nicht wieder beschrieben werden. Das elektrisch programmierbare ROM (EPROM) hingegen kann erneut beschrieben werden, das Löschen der Information ist aber nur mit UV-Licht möglich. Beim

elektrisch lösch- und programmierbaren ROM (EEPROM oder E²PROM) ist im Gegensatz zu EPROM-Speicherzellen ein Löschvorgang statt mit UV-Licht auch elektrisch möglich.

Die erste EEPROM-Zelle wurde 1978 von Perlegos *et al.* bei Intel entwickelt [PER78]. Die Abkürzung EEPROM setzt sich aus den Anfangsbuchstaben der Bezeichnung „**E**lectrically **E**rasable **P**rogrammable **R**ead **O**nly **M**emory“ zusammen. Das Löschen von Informationen ist innerhalb von wenigen Millisekunden bei EEPROM-Speichern wesentlich schneller als bei EPROMs aber deutlich langsamer als bei SRAM- oder DRAM-Speichern. Dafür sind EEPROMs „nicht-flüchtig“. Die Zeit, die die Information gespeichert werden kann, hängt aber stark von den Umgebungsbedingungen ab. Außerdem ist die Zahl der möglichen Schreib- und Löschvorgänge limitiert. Eine komplette EEPROM-Zelle besteht aus zwei Transistoren, dem Speicher- und dem Auswahltransistor.

Flash-EEPROM-Speicher funktionieren prinzipiell wie EEPROM-Zellen, sie benötigen aber deutlich weniger Fläche, da sie mit nur einem Transistor pro Bit auskommen. Dafür ist das Löschen bei ihnen immer nur gemeinsam für ein ganzes Segment eines Arrays möglich.

EEPROM-Speicher werden vor allem dort eingesetzt, wo Daten über einen längeren Zeitraum ohne Betriebsspannung verfügbar sein müssen und nur elektrisch beschrieben und ausgelesen werden können. Heute finden sie vor allem in der Computertechnik, im Kommunikationsbereich oder in der Automobilindustrie Verwendung.

4.1.2 Aufbau und Funktionsprinzip der EEPROM-Speicherzellen

Abbildung 4.1 zeigt zwei typische EEPROM-Zellen[1]. Bei beiden gibt es einen Source- und einen Drainbereich sowie zwei Gates, das so genannte Control-Gate und das Floating-Gate. Das Control-Gate kann von außen kontrolliert werden, das Potential des Floating-Gates kann hingegen nur über kapazitive Kopplung mit dem Control-Gate oder dem Drain gesteuert werden. Ist Ladung auf dem Floating-Gate gespeichert, so bleibt diese zunächst auch beim Abschalten der äußeren Spannung erhalten.

Die Zellvariante in Abbildung 4.1 a) ist eine Doppel-Poly-Zelle, denn sie besteht aus zwei Polysiliziumebenen, die das Control-Gate und das Floating-Gate bilden. Die Herstellung einer solchen Zelle ist sehr aufwendig und kostenintensiv, da eine zweite Polysiliziumebene benötigt wird. Eine Möglichkeit, die zweite Polysiliziumebene zu vermeiden, ist die Verwendung einer so genannten Single-Poly-Zelle, wie sie in Abbildung 4.1 b) dargestellt ist. Diese Variante beinhaltet nur eine Polysiliziumebene, die das Floating-Gate darstellt. Die Funktion des Control-Gates wird von einem dotierten Aktivgebietsbereich übernommen. Der Nachteil dieser Variante ist die größere benötigte Zellenfläche, d. h. bei hochintegrierten EEPROM-Schaltungen eignet sich die Single-Poly-EEPROM-Zelle weniger. Für die Fähigkeit der Speicherzelle, bei hohen Temperaturen genutzt zu werden, spielt dies aber keine Rolle. Die in dieser Arbeit verwendeten EEPROM-Zellen sind Single-Poly-Zellen (weitere Erklärungen dazu in Abschnitt 4.2.1).

In einem Array angeordnet benötigt jede EEPROM-Zelle einen zweiten Transistor, den so genannten Auswahl- oder Select-Transistor. Nur wenn dieser geöffnet ist, kann die eigentliche Speicherzelle angesteuert werden.

[1] Der Begriff „EEPROM-Zelle“ wird synonym für den Speichertransistor allein oder für die komplette Anordnung aus Speicher- und Auswahltransistor verwendet.

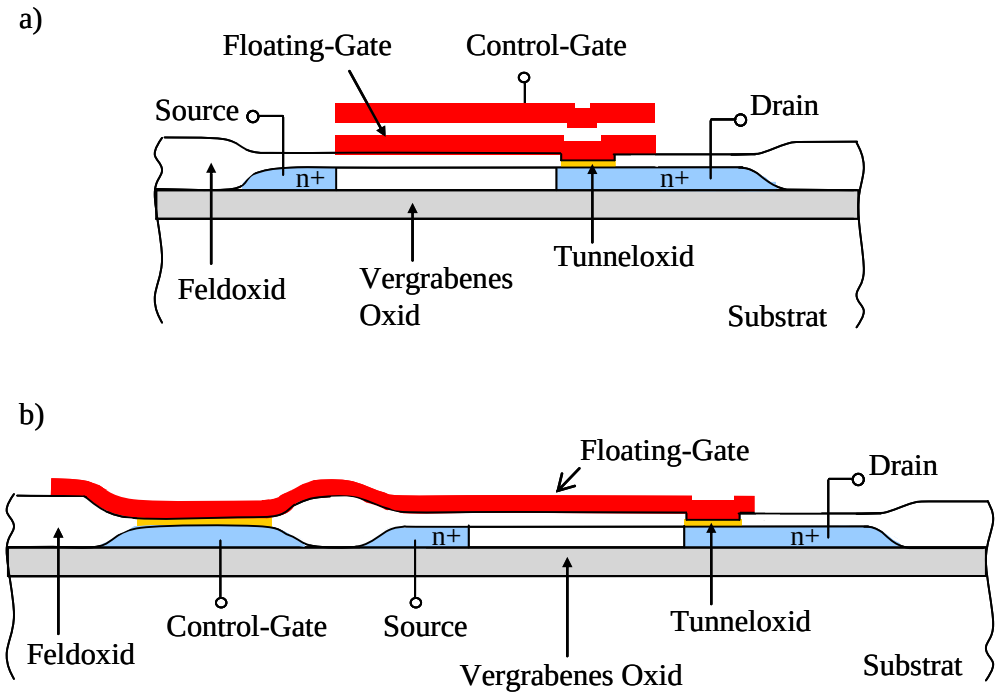


Abbildung 4.1: Schematische Darstellung typischer EEPROM-Zellen; a) als Doppel-Poly-Variante; b) als Single-Poly-Variante

Die binäre Information der Zelle (Zustand „1“ oder „0“) hängt von der auf dem Floating-Gate gespeicherten Ladung ab. Die Zelle zu programmieren oder zu löschen bedeutet, diese Ladung kontrolliert zu verändern. Das Programmieren und Löschen der EEPROM-Zellen basiert auf Fowler-Nordheim-Tunneln. Dafür wird an das Control-Gate bzw. an das Drain eine hohe Spannung (Programmierspannung[2]) angelegt und so ein elektrisches Feld aufgebaut, mit dessen Hilfe die Elektronen in einem Abschnitt zwischen Floating-Gate und Drain, wo die Oxiddicke reduziert ist, vom Drain-Gebiet zum Floating-Gate bzw. zurück tunneln können. Eine positive Spannung am Control-Gate vergrößert die Elektronenzahl auf dem Floating-Gate, eine positive Spannung am Drain vermindert sie. Der Fall, in dem ein hohes Potential am Control-Gate angelegt wird, wird „Löschen“ der Zelle genannt, der Fall des hohen Potentials am Drain wird entsprechend mit „Programmieren“ bezeichnet[3]. Die EEPROM-Zelle ist dann gelöscht bzw. programmiert.

Der Bereich reduzierter Oxiddicke von typischerweise 10 nm wird Tunneloxid genannt. Durch die geringe Oxiddicke ist es möglich, Fowler-Nordheim-Tunneln über dem Tunneloxid zu ermöglichen, den Gateoxidbereich aber nicht zu beeinflussen (siehe auch Kapitel 3). Andererseits wird durch eine Mindestdicke von 5 nm direktes Tunneln vermieden.

Die Ladung auf dem Floating-Gate bestimmt die Schwellenspannung U_{th} der EEPROM-Zelle, d.h. die Spannung, die an das Control-Gate angelegt werden muss, um den Kanal zwischen Source und Drain zu öffnen. Die Änderung der Schwellenspannung ΔU_{th} beim Programmieren bzw. Löschen hängt deshalb von der Menge und dem Vorzeichen der Ladung Q_F, die

[2] Auch beim Löschen spricht man von einer Programmierspannung.
[3] In der Literatur findet man häufig auch die umgekehrte Definition von Löschen und Programmieren, hier wird aber ausschließlich die im obigen Text definierte Variante verwendet.

sich auf dem Floating-Gate ansammelt und der Kapazität C_G zwischen dem Control-Gate und dem Floating-Gate ab [KOL86]:

$$\Delta U_{th} = -\frac{Q_F}{C_G} \qquad [4.1]$$

Durch die erhöhte Anzahl an Elektronen auf dem Floating-Gate im gelöschten Zustand liegt somit eine positive Schwellenspannungsänderung, d. h. eine größere Schwellenspannung als im neutralen Fall, vor. Die programmierte Zelle hat hingegen eine niedrigere Schwellenspannung als im neutralen Fall.

Die Differenz zwischen gelöschter und programmierter Schwelle wird Programmierfenster (engl. threshold voltage window) genannt. Die Größe dieses Fensters hängt von der Tunneloxiddicke, dem Koppelverhältnis (siehe Abschnitt 4.2.1), der angelegten Programmierspannung U_{pp} (siehe Abschnitt 4.2.4), der Temperatur und der Vorschädigung der Zelle ab. Ist das Fenster groß genug, d.h. kann der gelöschte und der programmierte Zustand nach festgelegten Kriterien eindeutig voneinander unterschieden werden, so hat die EEPROM-Zelle zwei logische Zustände, „1“ und „0“.

Um die Schwellenspannung einer EEPROM-Zelle zu lesen, gibt es zwei Möglichkeiten. Entweder kann ein bestimmter Strom eingeprägt und die zugehörige Control-Gate-Spannung gelesen werden, oder es wird der Strom in einem definierten Spannungsintervall gemessen und daraus die Schwellenspannung bestimmt. Für die zweite Methode wird am Control-Gate der Spannungsbereich durchfahren, in dem die Schwelle vermutet wird, während der Kanal zwischen Source und Drain geöffnet ist. Die Schwellenspannung entspricht dann der Control-Gate-Spannung bei einem festgelegten Kriterium, zum Beispiel bei einem Drainstrom von 1 µA. Die so genannte Lesespannung muss sich im Bereich der erwarteten Schwelle befinden, um die Ladung auf dem Floating-Gate nicht zu verschieben. Liegt beispielsweise die gelöschte Schwelle, auch „obere Schwelle“ genannt, bei 6 V, so lässt sie sich mit einer Lesespannung zwischen 4 V bis 8 V am Control-Gate ermitteln, ohne zusätzliche Elektronen auf das Floating-Gate zu ziehen oder sie wegzuschieben. Bei einer programmierten bzw. „unteren Schwelle“ von -2 V wäre eine Lesespannung von -4 V bis 0 V angemessen.

Eine Zelle, die weder gelöscht noch programmiert ist, besitzt eine „neutrale“ Schwelle. Diese muss nicht symmetrisch zwischen der oberen und der unteren Schwelle liegen, sondern hängt von der Geometrie der Zelle ab. Setzt man eine beschriebene EEPROM-Zelle für eine gewisse Zeit UV-Licht aus, so wird die Ladung auf dem Floating-Gate wieder in ihren ursprünglichen Status gebracht und man erhält die neutrale Schwelle.

Wie in der Einleitung angedeutet, gibt es bei EEPROM-Speichern zwei wesentliche Zuverlässigkeitsaspekte, den Erhalt der Ladung über einen definierten Zeitraum (Datenerhalt, engl. Data Retention) und die Fähigkeit der Zelle, mehrfach programmiert und gelöscht zu werden (Zyklenfestigkeit oder Datenwechselstabilität, engl. Endurance). In beiden Fällen wird der Ausfall der Zelle über die Größe des Programmierfensters, d.h. den Unterschied zwischen gelöschtem und programmiertem Zustand, definiert. Die physikalischen Hintergründe von Datenerhalt und Zyklenfestigkeit sowie die Vorgehensweise bei ihrer Untersuchung werden in den Abschnitten 4.1.3 und 4.1.4 erläutert. Ein Überblick über die beiden Zuverlässigkeitsuntersuchungen findet sich auch in den JEDEC-Standards JESD22-A117 und JESD22-A103C [JED22], [JED22b].

4.1.3 Datenerhalt (Data Retention)

Unter Datenerhalt versteht man die Fähigkeit einer EEPROM-Zelle, die gespeicherte Information über einen bestimmten Zeitraum auch ohne Anliegen einer Versorgungsspannung zu erhalten. Dies bedeutet, dass soviel der auf dem Floating-Gate befindlichen Ladung einer gelöschten oder programmierten EEPROM-Zelle erhalten bleiben muss, wie nötig ist, um die Zelle als programmiert bzw. gelöscht zu definieren.

Bei einem Überschuss oder Mangel an Elektronen auf dem Floating-Gate, wie es bei gelöschten bzw. programmierten Zellen der Fall ist, besteht im Bereich des Tunneloxids eine Potentialdifferenz zwischen dem Floating-Gate und dem Draingebiet. Sie ist groß genug, dass Elektronen weiterhin durch das dünne Oxid tunneln können (Fowler-Nordheim-Tunneln). Die Zelle verliert mit der Zeit an Ladung. Dieser Ladungsverlust ist stärker, je mehr Ladung auf dem Floating-Gate vorhanden ist, d.h. je weiter die gelöschte bzw. programmierte Schwelle vom neutralen Zustand entfernt ist. Sie hängt natürlich auch von der Dicke und der Qualität des Tunneloxids ab. Ein weiteres wichtiges Kriterium beim Datenerhalt ist die Temperatur. Da Fowler-Nordheim-Tunneln durch eine erhöhte Umgebungstemperatur beschleunigt wird, verliert die Zelle bei höherer Temperatur mehr Ladung.

Um den Datenerhalt, der während des Nutzungszeitraums der EEPROM-Speicher bei Betriebsbedingungen zu erwarten ist, zu simulieren, kann man beschleunigte Zuverlässigkeitstests durch Lagerungen bei hohen Temperaturen durchführen. Dazu werden EEPROM-Zellen in einen gelöschten bzw. programmierten Zustand gebracht und für eine gewisse Zeit ohne angelegte Spannung in einem Ofen gelagert. In definierten Zeitabständen werden die Schwellen bei Raumtemperatur ausgelesen. Der Verlauf der Schwellenspannungen über der Zeit wird notiert. Abbildung 4.2 zeigt einen typischen zeitlichen Verlauf der Schwellenspannungen programmierter und gelöschter Zellen bei einer Lagerungstemperatur von 250 °C.

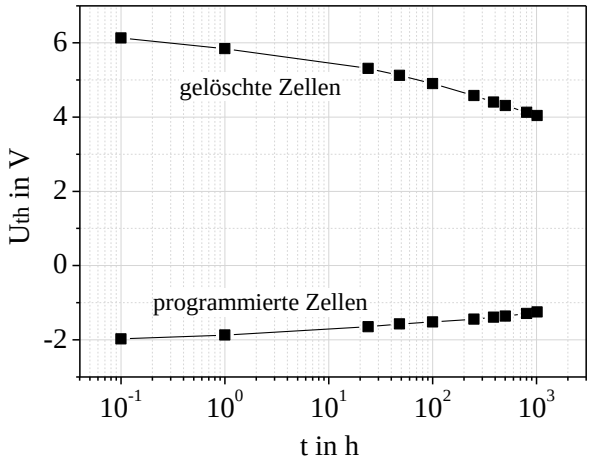


Abbildung 4.2: Beispiel für den Verlauf der Schwellen von gelöschten und programmierten EEPROM-Zellen in Abhängigkeit von der Zeit bei einer Lagerungstemperatur von 250 °C

Die Ausfallzeit der Zelle ist die Zeit, bei der die Schwellenspannung nicht mehr außerhalb eines festgelegten Programmierfensters liegt oder eine festgelegte prozentuale Änderung in Bezug auf das Fenster bei t = 0 h erfahren hat. Typische akzeptierte Änderungen des Programmierfensters liegen im Bereich von 10 % [DeS99] bis 40 % [PAP95].

Je länger die Lagerungszeit und je höher die Lagerungstemperatur sind, desto mehr Ladung geht verloren und beide Schwellen nähern sich der neutralen Schwelle an, bis sie am Ende nicht mehr voneinander zu unterscheiden sind. Dabei ist die Abnahme der oberen Schwelle nicht immer symmetrisch zur Zunahme der unteren Schwelle. Aus diesem Grund wird meistens die Änderung des Programmierfensters betrachtet, um das Verhalten verschiedener Zellen zu vergleichen.

Wenn die Lagerung bei mehreren Temperaturen, für die jeweils eine Ausfallzeit gemessen werden kann, durchgeführt wird, lässt sich mit Hilfe des Arrheniusgesetzes ein Arrheniusgraph erstellen und daraus eine Aktivierungsenergie für den Ladungsverlustmechanismus ermitteln (siehe auch Kapitel 2). Mit dieser Aktivierungsenergie kann man dann auf die Ausfallzeit bei der Einsatztemperatur extrapolieren [CAR98]:

$$t_{\text{fail}} = t_0 \cdot \exp\left(\frac{E_A}{k \cdot T}\right) \qquad [4.2]$$

Hierbei ist t_{fail} die Zeit bis zum Ausfall der Zelle, t_0 eine Zeitkonstante, E_A die Aktivierungsenergie, k die Boltzmann-Konstante und T die Einsatztemperatur in Kelvin.

Die Lebensdauer für EEPROM-Speicher bei Raumtemperatur beträgt je nach Anwendungsgebiet 10 bis 20 Jahre. Bei einer Einsatztemperatur von 250 °C, wie im H10-Prozess vorgesehen, ist die Lebensdauer einer EEPROM-Zelle deutlich geringer als bei Raumtemperatur. Für die EEPROM-Zellen des H10-Prozesses wird eine Einsatzdauer von mindestens 1000 h bei 250 °C angestrebt.

Beschleunigte Lagerungen werden typischerweise bei Temperaturen um die 250 °C durchgeführt [HOU08]. Im H10-Prozess ist 250 °C aber schon die Betriebstemperatur, so dass wie auch schon bei den Experimenten zur Zuverlässigkeit des Gateoxids noch höhere Temperaturen nötig sind, um die Tests zu beschleunigen. Für die Untersuchungen zum Datenerhalt der H10-EEPROM-Zellen wurden deshalb EEPROM-Zellen bei Temperaturen bis 450 °C gelagert. Um auch den Vergleich zu den sonst üblichen Lagerungstemperaturen zu gewährleisten, wurden auch Lagerungen bei 160 °C und 250 °C durchgeführt. Dabei sind sowohl die hohen Lagerungstemperaturen an sich als auch der insgesamt sehr weite Temperaturbereich von 160 °C bis 450 °C, der die Einsatztemperatur sogar mit einschließt, hervorzuheben. Für die Datenanalyse gilt dann die Frage, ob die bekannten Beschleunigungsformeln anwendbar sind und/oder ob neue Ausfallmechanismen eine Rolle spielen.

4.1.4 Zyklenfestigkeit (Endurance)

Die Zyklenfestigkeit, auch Datenwechselstabilität genannt, bezeichnet die Fähigkeit einer EEPROM-Zelle, mehrfach beschrieben und wieder gelöscht werden zu können. Ein kombinierter Schreib- und Löschvorgang wird als Zyklus bezeichnet und das abwechselnde Programmieren und Löschen nennt sich Zykeln.

Im Gegensatz zum Datenerhalt, bei dem das Oxid nicht geschädigt wird, sondern Elektronen durch Tunneln verloren gehen, ist bei der Zyklenfestigkeit die Reduzierung des Programmierfensters eine Folge der Oxiddegradation. Mit zunehmender Zahl an Schreib- und Löschvorgängen, bei denen jedes Mal Ladung durch das Tunneloxid fließt, nimmt auch die Anzahl der Defekte und damit die Degradation des Isolators zu. Durch die zunehmende Oxidschädigung, die sich in der Besetzung bzw. Generierung von Haftstellen äußert, ändern sich

das über dem Oxid anliegende Potential und dadurch auch der Stromfluss (siehe auch Kapitel 3). Als Folge verändern sich die Schwellenspannungen und damit auch das Programmierfenster. Die Zyklenfestigkeit wird deshalb als Veränderung des Programmierfensters mit der Anzahl der Zyklen beschrieben.

In der praktischen Durchführung wird eine EEPROM-Zelle bei einer bestimmten Temperatur gezykelt und die Schwelle in definierten Zyklenabständen gelesen. Abbildung 4.3 gibt ein Beispiel für eine so genannte Endurance-Kurve, die die Änderung der gelöschten und der programmierten Schwelle einer Zelle in Abhängigkeit von der Anzahl der Zyklen zeigt. Dafür wurde eine EEPROM-Zelle bei 25 °C mehr als eine Million Mal gezykelt.

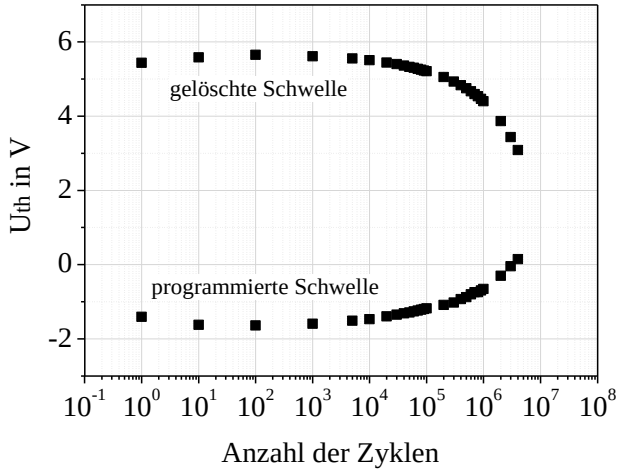


Abbildung 4.3: Beispiel für den Verlauf der Schwellen von gelöschten und programmierten EEPROM-Zellen in Abhängigkeit von der Anzahl der Zyklen bei einer Temperatur von 25 °C

Die Veränderung der Schwellenspannung mit der Anzahl der Zyklen kann in zwei Phasen eingeteilt werden [EUZ81]. In der ersten Phase nach nur wenigen Zyklen weitet sich das Programmierfenster auf. Der Grund dafür sind positive Ladungen, die sich am Ende des Schreibvorgangs an der Siliziumdioxid-Silizium-Grenzfläche befinden. Sie verkleinern die Tunnelbarriere, wodurch der Fowler-Nordheim-Strom steigt und damit das Programmierfenster größer wird. Mit zunehmender Zyklenzahl tritt dann die zweite Phase ein, in der vor allem Elektronen in das Tunneloxid injiziert werden. Diese negativen Ladungen reduzieren das elektrische Feld über dem Tunneloxid, der Tunnelstrom nimmt ab und das Programmierfenster wird kleiner. Dabei ist wie auch beim Datenerhalt die Veränderung des Fensters nicht in jedem Fall symmetrisch.

Die Anzahl der möglichen Lösch- und Schreibvorgänge der EEPROM-Zellen, die von den Herstellern für die Anwendung bei Raumtemperatur garantiert werden, beträgt typischerweise bis zu 100.000 Zyklen. Für die H10-EEPROM-Zellen ist eine Anzahl von mindestens 1000 Zyklen bei 250 °C das Ziel.

Die Untersuchungen zur Zyklenfestigkeit wurden bei verschiedenen Programmierspannungen an Zellen unterschiedlicher Tunneloxiddicken in einem Temperaturbereich von -40 °C bis 400 °C durchgeführt. Da bisher keine Extrapolationsmodelle für Endurance-Untersuchungen bei hohen Temperaturen hin zu Betriebstemperaturen existieren, wird die Zyklenfestigkeit in der Regel nur bei der maximalen Betriebstemperatur untersucht, d. h. für Nicht-Hochtempe-

ratur-Prozesse bis 125 °C oder maximal 175 °C. Das Ziel dieser Arbeit ist, eine Abhängigkeit der Zyklenzahl von der Temperatur und dem elektrischen Feld abzuleiten.

4.2 Beschreibung der untersuchten EEPROM-Speicherzellen

4.2.1 Single-Poly-EEPROM-Einzelzellen

Die in den Untersuchungen zu dieser Arbeit verwendeten Single-Poly-EEPROM-Einzelzellen wurden in früheren Arbeiten am Fraunhofer IMS entwickelt und sind in zwei Doktorarbeiten ausführlich beschrieben [WAL96], [GOG97]. Abbildung 4.4 zeigt in einer schematischen Darstellung den Aufbau der untersuchten EEPROM-Speicherzellen.

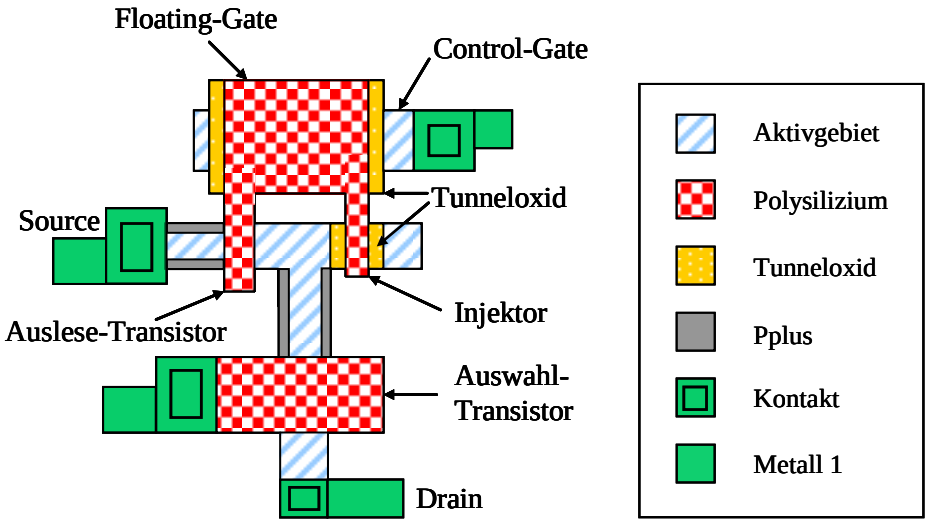


Abbildung 4.4: Schema der untersuchten Single-Poly-EEPROM-Einzelzelle

Die eigentliche EEPROM-Zelle ist im oberen Bereich dargestellt. Es gibt zwei Tunneloxidbereiche, zwischen dem Control-Gate und dem Floating-Gate und unter dem Injektor, die die gleiche Dicke aber eine unterschiedliche Fläche haben. Das Tunneloxid wird in einem RTA-Schritt (Rapid Thermal Annealing) unter Sauerstoffatmosphäre gefertigt. Der Siliziumbereich unter dem Tunneloxid ist mit Phosphor n-dotiert. Als Standard-Tunneloxiddicke für die H10-EEPROM-Zellen werden 11 nm festgelegt. Weil unter dem Injektorfinger die Tunneloxidfläche viel kleiner ist als im oberen Bereich des Control-Gates, liegt gemäß den Gesetzmäßigkeiten für kapazitive Spannungsteiler zwischen dem Injektor und dem Aktivgebietsbereich der größte Teil der Programmierspannung an. Dadurch können beim Anlegen der Programmier- und Löschspannungen an Source, Drain und Control-Gate die Elektronen im Bereich des Injektors auf das Floating-Gate bzw. vom Floating-Gate herunter tunneln. Das Verhältnis der beiden Tunneloxidflächen wird Koppelverhältnis genannt. Die Größe dieses Verhältnisses bestimmt, wie viele Elektronen bei einer bestimmten Spannung tunneln können. Damit beeinflusst es auch die Lage der gelöschten und der programmierten Schwellen sowie die Größe des Programmierfensters. Bei den verwendeten EEPROM-Zellen beträgt das Koppelverhältnis circa 9.

Getrennt vom Injektorfinger stellt ein zweiter Polysiliziumsteg den Auslesetransistor dar, über den die Schwellenspannungen gemessen werden. Das Polysilizium-Gate des Auswahl-Transistors wird Select-Gate genannt. Es trennt den Drain-Bereich der EEPROM-Zelle vom Rest der Speicherzelle. Nur wenn der Auswahl-Transistor geöffnet ist, kann das Drain angesteuert werden. Für eine Einzelzelle ist die Anwesenheit eines Auswahl-Transistors nicht nötig, in einem Array aber ermöglicht er, eine bestimmte EEPROM-Zelle auszuwählen bzw.

alle anderen Zellen zu sperren. Alle Transistoren der Single-Poly-EEPROM-Zellen sind NMOS-Transistoren.

Für eine größere Statistik an Messdaten und damit quantitativ bessere Abschätzungen bieten sich Untersuchungen an in Arrays formierten Speicherzellen an. Ein großer Teil der Experimente zu dieser Arbeit wurde aber an Single-Poly-Einzelzellen durchgeführt, da zu Beginn der Zuverlässigkeitsuntersuchungen noch keine EEPROM-Arrays vorlagen.

4.2.2 EEPROM-Arrays

Ein H10-EEPROM-Array besteht aus einer regelmäßigen Anordnung von insgesamt 384 Single-Poly-Einzelzellen. Die Zellen sind in acht Blöcken zu je sechs Zeilen und acht Spalten angeordnet. Jeder Block besitzt ein gemeinsames Control-Gate, so wie jede Zeile ein gemeinsames Select-Gate und jede Spalte ein gemeinsames Drain hat. Die acht Control-Gates, sechs Select-Gates und acht Drains sind alle separat an insgesamt 22 Pads angeschlossen. Jede Einzelzelle im Array kann somit über eine andere Pad-Kombination angesteuert werden. Die Source-Anschlüsse aller 384 Einzelzellen laufen auf ein gemeinsames Pad.

Das Programmieren und Löschen der Zellen im Array erfolgt über die Auswahl des Blocks (Control-Gate), der Zeile (Select-Gate) und der Spalte (Drain). Wird ein EEPROM in einem bestimmten Block programmiert oder gelöscht und ist damit selektiert, müssen die anderen Blöcke jeweils blockiert (deselektiert) werden, damit deren Zellen nicht beeinflusst werden und keine so genannten „Disturb-Effekte“ [4] auftreten. Gleiches gilt für die Zeilen und Spalten.

Um einen Überblick über die Potentialverhältnisse an den Anschlüssen der EEPROM-Einzelzellen und der EEPROM-Arrays zu bekommen, wenn diese für die Experimente zum Datenerhalt vorbereitet oder auf ihre Zyklenfestigkeit hin überprüft werden, werden im nächsten Abschnitt die Einstellungen für das Programmieren, Löschen und Lesen der Zellen erläutert.

4.2.3 Programmieren, Löschen und Auslesen von Einzelzellen und Arrays

Das Programmieren, Löschen und Auslesen der EEPROM-Zellen erfolgt bei Raumtemperatur mit Hilfe von LabVIEW-Programmen, die den Pulsgenerator und den Parameteranalyser ansteuern können (siehe Kapitel 2). Die dabei an den jeweiligen Anschlüssen anliegenden Spannungen sind in Tabelle 4.1 zusammengefasst.

Während des Programmierens (Löschens) wird die Programmierspannung U_{pp} für wenige Millisekunden an das Drain (Control-Gate) angelegt. Für die H10-EEPROM-Zellen wurden 8 ms als Standardpulsdauer festgelegt.

Da es sich um EEPROM-Zellen auf SOI-Substrat handelt, liegt der Waferchuck beim Programmieren, Löschen und Lesen immer auf einem Potential von 0 V. Das Potential am Select-Gate entspricht beim Programmieren immer dem Potential am Drain und beim Löschen dem Potential am Control-Gate. Beim Lesen wird der Auswahl-Transistor durch ein Potential von 5 V am Select-Gate geöffnet. Im Fall der Arrays ist darauf zu achten, dass die

[4] Unter einem „Disturb-Effekt“ versteht man die Beeinflussung der Schwellenspannung einer deselektierten Zelle durch das Programmieren, Löschen oder Lesen einer anderen Zelle [HOU08].

nicht ausgewählten Zellen, d h. die deselektierten Select-Gates, ein Potential von 0 V sehen, damit ihre Auswahl-Transistoren sperren.

	Programmieren	**Löschen**	**Auslesen $U_{th,pr}$**	**Auslesen $U_{th,ge}$**
$\boldsymbol{U_{CG}}$	0 V	U_{pp} (z. B. 16 V)	je nach U_{pp} ca. -6 bis 2 V	je nach U_{pp} ca. 2 bis 10 V
$\boldsymbol{U_{D}}$	U_{pp} (z. B. 16 V)	0 V	0,1 V	0,1 V
$\boldsymbol{U_{SG}}$	U_{pp} (z. B. 16 V)	U_{pp} (z. B. 16 V)	5 V	5 V
$\boldsymbol{U_{S}}$	floatend	floatend	0 V	0 V
$\boldsymbol{U_{Chuck}}$	0 V	0 V	0 V	0 V

Tabelle 4.1: Einstellungen am Control-Gate (CG), am Drain (D), am Select-Gate (SG), an der Source (S) und am Chuck beim Programmieren und Löschen von EEPROM-Einzelzellen sowie beim Auslesen der programmierten Schwellen $U_{th,pr}$ und der gelöschten Schwellen $U_{th,ge}$

Wird die Zelle programmiert, werden 0 V an das Control-Gate angelegt, entsprechend 0 V an das Drain beim Löschen, wobei der Source-Kontakt in beiden Fällen floatend ist. Bei der Messung wird der Kanal des Auslesetransistors durch 0,1 V am Drain und 0 V an der Source geöffnet und das Auslesen der Schwellenspannungen ermöglicht. Dazu wird am Control-Gate ein begrenzter Spannungsbereich durchfahren, in dem die Schwelle vermutet wird (siehe auch Abschnitt 4.2.1). Dieser Bereich muss je nach Programmierspannung angepasst werden. Die Schwellenspannungen werden als Control-Gate-Spannung bei einem Drainstrom von 1 µA bestimmt. Für die korrekte Schwellenspannungsbestimmung bei den EEPROM-Arrays ist es wichtig, dass die nicht ausgewählten Control-Gate-Blöcke durch ein negatives Potential am Control-Gate blockiert werden. Dieses muss zudem unterhalb der Schwelle liegen, die gelesen wird.

Um eine stabile Schwelle zu erhalten, werden EEPROM-Zellen, die für Experimente zum Datenerhalt gelagert werden, zehnmal abwechselnd programmiert und gelöscht und am Ende im programmierten bzw. im gelöschten Zustand belassen. Dieses zehnmalige Zykeln wird „Initialisieren“ genannt. Verändert sich die Schwelle einer Zelle durch Temperaturlagerung oder durch häufiges Zykeln, muss eventuell der beim Auslesen abgesuchte Control-Gate-Bereich angepasst werden.

Für die Untersuchungen zum Datenerhalt wurden nur EEPROM-Einzelzellen und EEPROM-Arrays auf Waferebene verwendet. Die Endurance-Messungen bei Temperaturen bis 400 °C wurden teilweise auch an EEPROM-Zellen im Gehäuse vorgenommen.

4.2.4 Programmierbarkeit der EEPROM-Speicherzellen

Um für die Untersuchungen zum Datenerhalt und zur Zyklenfestigkeit geeignete Programmierspannungen zu finden, wurden EEPROM-Einzelzellen auf ihre Programmier- und Löschbarkeit bei verschiedenen Programmierspannungen hin untersucht. Abbildung 4.5 zeigt die jeweiligen Schwellenspannungen der gelöschten und der programmierten Schwellen bei 25 °C (a) und 250 °C (b) für zwei verschiedene Tunneloxiddicken. Das Potential am Select-Gate entspricht beim Programmieren dem Potential am Drain und beim Löschen dem Potential am Control-Gate.

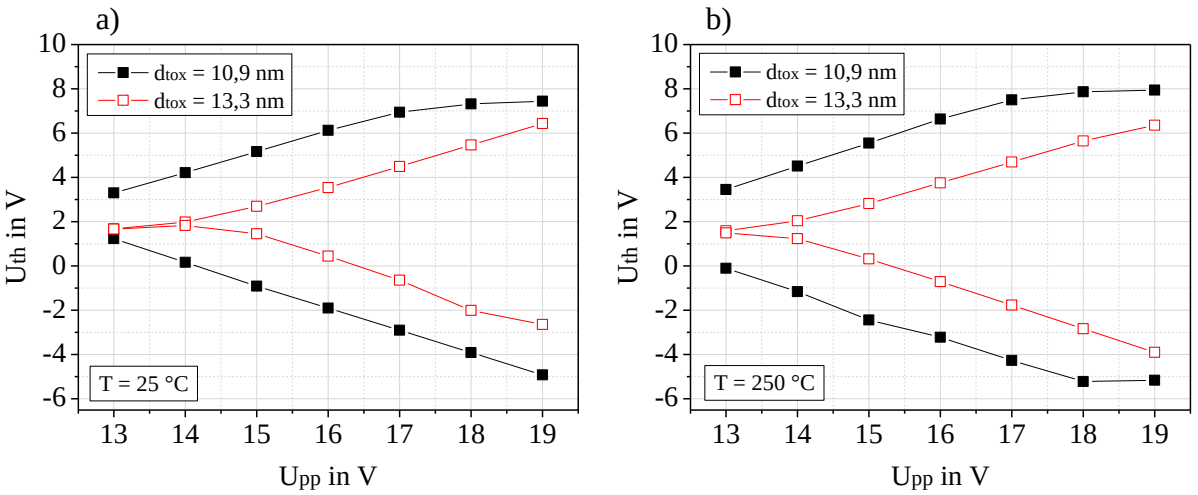


Abbildung 4.5: Schwellenspannungen von gelöschten und programmierten EEPROM-Zellen in Abhängigkeit von der Programmierspannung U_{pp} bei einer Temperatur von 25 °C (a) bzw. 250 °C (b) für Tunneloxiddicken d_{tox} von 10,9 nm (ausgefüllte Kästchen) und 13,3 nm (leere Kästchen)

Mit zunehmender Programmierspannung nehmen die untere Schwelle ab und die obere Schwelle zu, d. h. das Programmierfenster weitet sich. Bei 250 °C ist es bei gleichen Potentialen immer etwas größer als bei 25 °C, weil der Tunnelstrom mit der Temperatur zunimmt. Während für das dickere Tunneloxid eine Programmierspannung von 13 V nicht ausreicht, um die Schwellen voneinander zu trennen, erreicht man für das dünnere Tunneloxid bei U_{pp} = 19 V eine gewisse Sättigung. Die Sättigung kommt dadurch zustande, dass mit zunehmender Elektronenzahl auf dem Floating-Gate auch die Zahl der Elektronen, die vom Floating-Gate zum Control-Gate tunneln, zunimmt. Ab einer gewissen Programmierspannung herrscht dann ein Gleichgewicht zwischen der Zahl der Elektronen, die vom Injektor weiter auf das Floating-Gate gebracht wird, und der Zahl der Elektronen, die zum Control-Gate tunneln, so dass insgesamt die Ladung auf dem Floating-Gate und damit die Schwelle stabil bleibt.

Bei der mittleren Programmierspannung von 16 V ist das Fenster für die Standard-Tunneloxiddicke von circa 11 nm bei 25 °C etwa 8 V und bei 250 °C fast 10 V groß. Dies ist groß genug, um die gelöschte und die programmierte Schwelle voneinander zu unterscheiden, dennoch weit genug vom Sättigungsbereich entfernt und bei einer begrenzten Anzahl an Pulsen auch nicht schädigend für das Tunneloxid. Deshalb werden 16 V als Standard-Programmierspannung ausgewählt.

Alle EEPROM-Zellen, die auf ihren Ladungserhalt hin überprüft werden, wurden vorher mit einer Programmierspannung von 16 V initialisiert. Für die Untersuchungen zur Zyklenfestigkeit werden Programmierspannungen von 13 V bis 19 V getestet.

4.3 Untersuchungen zur Zuverlässigkeit der EEPROM-Speicherzellen

4.3.1 Datenerhalt (Data Retention)

Wie in Abschnitt 4.1.3 erläutert, ist die Bestimmung des Datenerhalts und damit der Dauer des Erhalts der gespeicherten Informationen ein essentieller Zuverlässigkeitsaspekt von EEPROM-Zellen. Um diesen Aspekt genauer zu untersuchen, wurden auf sechs Waferstücken Einzelzellen bei Raumtemperatur initialisiert und dann die Hälfte der Zellen im gelöschten, die andere Hälfte im programmierten Zustand bei Temperaturen zwischen 160 °C und 450 °C gelagert[5]. Nach gewissen Zeitabständen wurden die Wafer aus dem Ofen genommen und die Schwellen wieder bei Raumtemperatur ausgelesen. Abbildung 4.6 a) zeigt die zeitliche Entwicklung der gelöschten und der programmierten Zellen für die verschiedenen Temperaturen (siehe auch [GRE11]). In Abbildung 4.6 b) ist das Programmierfenster in Abhängigkeit von der Lagerungszeit aufgetragen. Aufgrund der logarithmischen *x*-Achsendarstellung sind die Werte der Schwellenspannungen von t = 0 h bei t = 0,1 h eingezeichnet. Die Tunneloxiddicke beträgt etwa 12 nm.

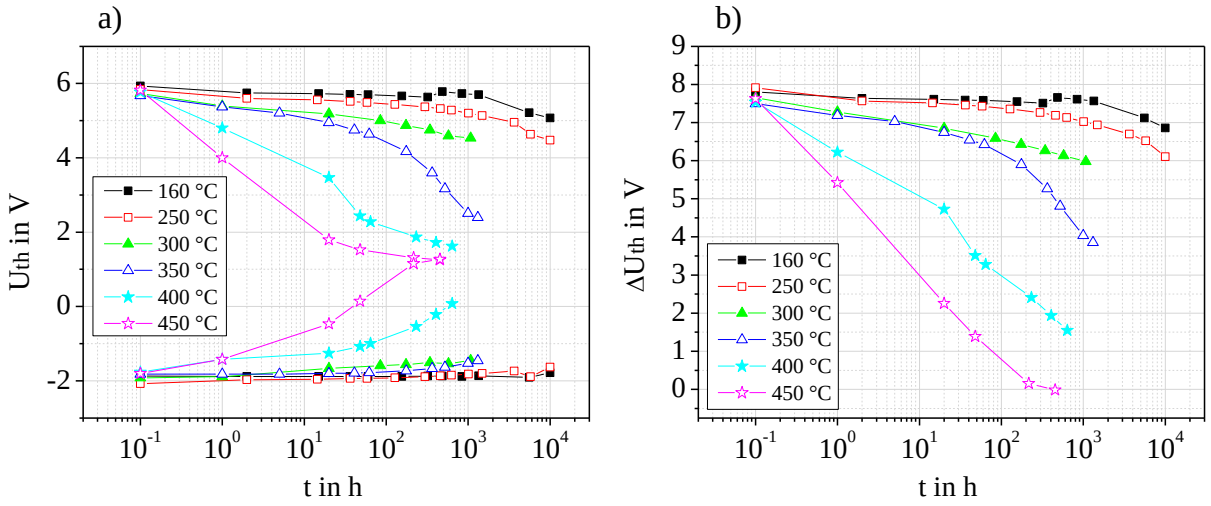


Abbildung 4.6: Schwellenspannungen für gelöschte (obere Schwellen) und programmierte (untere Schwellen) EEPROM-Einzelzellen in Abhängigkeit von der Zeit bei Lagerungstemperaturen zwischen 160 °C und 450 °C (a); Differenz der oberen und der unteren Schwellen (Programmierfenster ΔU_{th}) in Abhängigkeit von der Zeit (b); die Messwerte bei t = 0 h sind wegen der logarithmischen *x*-Achsendarstellung bei t = 0,1 h eingezeichnet

In Abbildung 4.6 a) ist zu erkennen, dass die oberen Schwellen vor Beginn der Lagerung bei etwa 6 V und die unteren Schwellen bei etwa -2 V liegen. Die gelöschten Schwellen sinken mit der Zeit, weil die sich auf dem Floating Gate befindliche Ladung abnimmt. Die unteren Schwellen bleiben im selben Zeitraum bei 160 °C bis 350 °C nahezu konstant und steigen für 400 °C und 450 °C an, jedoch langsamer, als die jeweiligen oberen Schwellen abfallen. Je größer die Temperatur ist, desto stärker bzw. schneller verläuft die Änderung. Die Kurvenform ist zudem für die verschiedenen Temperaturen unterschiedlich. Bis zu einer Temperatur von 350 °C scheint das Verhalten der oberen Schwellen im ersten Zeitabschnitt logarithmisch zu sein, ab einem bestimmten Punkt aber sinkt die Schwelle stark ab. Bei 400 °C und 450 °C liegt zunächst eine starke Veränderung der Schwellen vor, die aber mit der

[5] Die Lagerung bei 450 °C ist nur deshalb möglich, weil die in diesem Abschnitt verwendeten Wafer eine Aluminium- statt einer Wolframmetallisierung erhalten haben, d. h. die Oxidation der Pads weniger kritisch ist (vergleiche auch Kapitel 5).

Zeit schwächer wird. Nach mehr als 100 h bei 450 °C treffen sich die beiden Schwellen der H10-EEPROMs bei etwa 1,6 V. Dieser Wert entspricht der neutralen Schwelle von unbeschriebenen EEPROM-Zellen, die bei der Löschung programmierter Zellen mit UV-Licht erreicht werden würde. Die neutrale Schwelle liegt etwas unterhalb der Mitte zwischen der gelöschten und der programmierten Anfangsschwelle von 6 V bzw. -2 V.

Der in Abbildung 4.6 b) dargestellte Verlauf des Programmierfensters wird bis einschließlich 350 °C von der oberen Schwelle dominiert, bei 400 °C und 450 °C fließt auch die Zunahme der unteren Schwelle mit ein. Bei t = 0 h beträgt das Fenster zwischen 7,5 V und 8 V. Nach 10.000 h bei 160 °C hat das Programmierfenster um etwa 1 V, nach 10.000 h bei 250 °C um etwa 2 V abgenommen.

Dass der zeitliche Verlauf des Programmierfensters nur in Abschnitten logarithmisch ist, mit der Temperatur variiert und für die untere und die obere Schwelle unterschiedlich ist, ist auch in der Literatur zu finden [DeS99], [DeS99b], [KIM03]. Um dieses Verhalten zu erklären, sind weitere Analysen notwendig. Als erster Schritt wird dazu in Abbildung 4.6 der zugehörige Arrheniusgraph betrachtet.

Arrheniusgraph

Wie in Abschnitt 4.1.3 erläutert, ist für die Berechnung der Aktivierungsenergie nach Arrhenius die Festlegung einer Ausfallzeit t_{fail} nötig, die meistens bei einer bestimmten prozentualen Änderung des Programmierfensters definiert wird. Da der Verlauf der Kurven in Abbildung 4.6 von der Temperatur abhängt und sich mit der Zeit ändert, spielt die Wahl dieses Kriteriums eine entscheidende Rolle, was in der Literatur oft nicht beachtet wird. In Abbildung 4.7 sind deshalb in einem Arrheniusgraphen die Logarithmen der Ausfallzeiten t_{fail} bei 15 % und bei 25 % Abnahme des Programmierfensters in Abhängigkeit von 1/(kT) aufgetragen. Die Steigungen der Fit-Kurven entsprechen den Aktivierungsenergien.

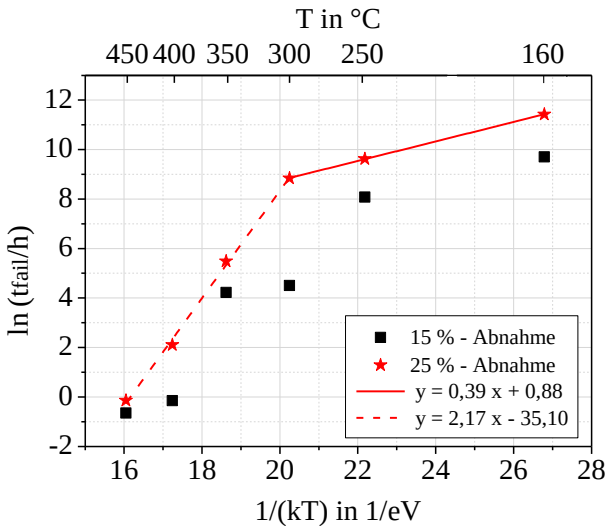


Abbildung 4.7: Logarithmus der Ausfallzeiten bei 15 % (Kästchen) bzw. 25 % (Sterne) Abnahme des Programmierfensters in Abhängigkeit von 1/(kT); die Messpunkte wurden aus Abbildung 4.6 b) gewonnen

Für die Interpretation des Arrheniusgraphen sind mehrere Punkte von Bedeutung. Aus Abbildung 4.7 kann man entnehmen, dass im Bereich zwischen 160 °C und 450 °C scheinbar

nicht nur eine Aktivierungsenergie existiert, sondern der Graph in zwei Bereiche eingeteilt werden muss. Bezieht man sich auf die Ausfallzeiten nach 25 % Änderung des Programmierfensters, gibt es zwei Abschnitte mit unterschiedlicher Steigung, d. h. unterschiedlicher Aktivierungsenergie. Zwischen 160 °C und 300 °C ist die Aktivierungsenergie mit etwa 0,4 eV deutlich geringer als im Abschnitt zwischen 300 °C und 450 °C, wo sie ungefähr 2,2 eV beträgt. Bezieht man den Datenpunkt bei 250 °C mit in den Abschnitt mit der größeren Steigung mit ein, erhält man eine Aktivierungsenergie von etwa 1,7 eV.

Die Daten des Arrheniusgraphen bei einem Bezug auf 15 % Änderung des anfänglichen Programmierfensters schwanken stark. Bei den höheren Temperaturen von 400 °C und 450 °C muss berücksichtigt werden, dass die erste Messung nach der Lagerung erst nach einer Stunde vorgenommen wurde, 15 % Änderung aber schon nach weniger als einer Stunde erreicht sind[6]. Aus diesem Grund muss die Ausfallzeit bei diesen Temperaturen mit einer Interpolation ermittelt werden, wodurch Ungenauigkeiten bei den beiden Messpunkten möglich sind. Zusätzlich besteht bei kleineren Lagerungszeiten auch eine größere Unsicherheit in der Exaktheit der Lagerungszeit.

Der Grund für die Unterschiede zwischen den Daten der 15 % - und der 25 % - Änderung des Programmierfensters liegt hauptsächlich im bereits erwähnten unterschiedlichen zeitlichen Verlauf bei den verschiedenen Temperaturen. Während sich eine Änderung von 15 % bei einer Lagerungstemperatur von 350 °C noch im Bereich des schwächeren Abfalls der oberen Schwelle befindet, ist dies beispielsweise bei 250 °C erst im stärker abfallenden Anteil der Fall. Eine Änderung von 25 % hingegen erreichen alle sechs Kurven erst im stark abfallenden Anteil, wodurch eine größere Vergleichbarkeit der Temperaturen gegeben ist.

Temperaturabhängige Aktivierungsenergie

Statt einer Aufteilung des Arrheniusgraphen in zwei Bereiche mit unterschiedlichen Steigungen wäre eine andere mögliche Interpretation des Arrheniusgraphen, dass die Arrheniusgleichung für den Datenerhalt von EEPROM-Speicherzellen nicht anwendbar ist, da sich keine Aktivierungsenergie im Sinne von Arrhenius ermitteln lässt. De Salvo *et al.* betrachten die Möglichkeit, dass die Aktivierungsenergie temperaturabhängig ist, d. h. dass für jede Temperatur eine andere Aktivierungsenergie gilt und leiten daraus das so genannte T-Modell ab [DeS99], [DeS99b]. Darin ist der Logarithmus der Ausfallzeiten proportional zur Temperatur. Mit den von De Salvo *et al.* ermittelten Konstanten würde sich bei uns für 160 °C eine Aktivierungsenergie von etwa 0,7 eV und für 450 °C eine Aktivierungsenergie von mehr als 2 eV ergeben.

Die Anwendung des T-Modells ist in Abbildung 4.8 dargestellt. Würde dieses Modell auf den hier vorliegenden Fall zutreffen, müssten die aufgetragenen Punkte eine Gerade bilden. Dies ist nicht eindeutig der Fall, aber durch die genannten Ungenauigkeiten bei der Bestimmung der Ausfallzeiten auch nicht auszuschließen. Allerdings ist auch in den Daten von De Salvo *et al.* die Möglichkeit gegeben, den Arrheniusgraphen in zwei Bereiche mit unterschiedlichen Aktivierungsenergien aufzuteilen, statt jeder Temperatur eine eigene Aktivierungsenergie zuzuordnen.

[6] Da vor der Lagerung nicht bekannt war, wie schnell die Werte abnehmen würden, wurden die Zellen erst nach 1 h Lagerungszeit ausgelesen. Wie später in diesem Kapitel deutlich wird, kann eine Lagerung auf demselben Wafer aber nicht wiederholt und auch nicht mit einem anderen Wafer kombiniert werden, so dass keine Möglichkeit bestand, Messpunkte zwischen $t = 0$ h und $t = 1$ h aufzunehmen.

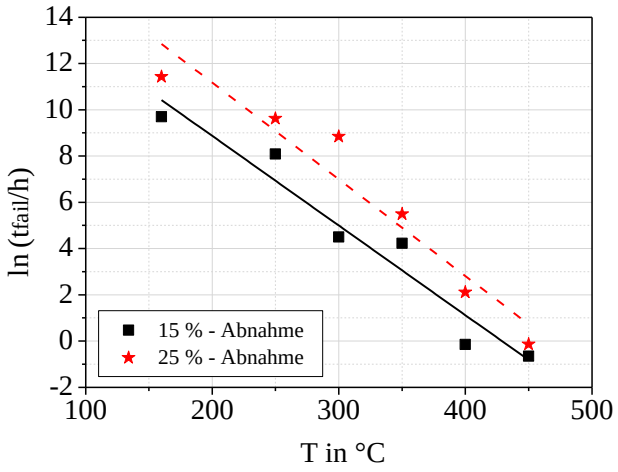


Abbildung 4.8: Logarithmus der Ausfallzeiten bei 15 % (Kästchen) bzw. 25 % (Sterne) Abnahme des Programmierfensters in Abhängigkeit von *T*; die Messpunkte wurden aus Abbildung 4.6 b) gewonnen

Man kann festhalten, dass sich aufgrund der bei den verschiedenen Temperaturen unterschiedlich verlaufenden zeitlichen Entwicklung des Programmierfensters Abweichungen für verschiedene prozentuale Abnahmen ergeben können. Dies bedeutet, dass die Definition des Ausfallkriteriums zu unterschiedlichen Ergebnissen führen kann. Dadurch ist es schwierig, korrekte Aktivierungsenergien zu ermitteln bzw. überhaupt festzustellen, ob die Arrheniusgleichung anwendbar ist. Die Einteilung von Abbildung 4.7 in zwei Bereiche mit unterschiedlichen Aktivierungsenergien ist eine mögliche Interpretation, eine temperaturabhängige Aktivierungsenergie eine andere.

Um diese Beobachtungen genauer zu analysieren, werden im nächsten Abschnitt die Ergebnisse der Lagerungen der H10-EEPROMs mit Untersuchungen an EEPROM-Zellen einer anderen Technologie verglichen, um herauszufinden, ob die Beobachtungen bei Lagerungstemperaturen oberhalb von 250 °C auch dort zutreffen. Die untersuchten Zellen sind zudem auf einem anderen Substrat gefertigt worden (Bulk statt SOI) und haben damit auch ein anderes Zelllayout. Zudem ist das Tunneloxid dünner (ca. 9,6 nm).

Vergleich mit EEPROM-Zellen auf Bulk-Substraten

In den Abbildungen 4.9 a) und b) ist das Verhalten der Schwellen gelöschter und programmierter EEPROM-Einzelzellen auf Bulk-Wafern in Abhängigkeit von der Zeit dargestellt. Auch hier wurden die Zellen bei Raumtemperatur initialisiert und gelesen.

Im Gegensatz zu den Beobachtungen bei den H10-EEPROM-Zellen ist bei den Bulk-EEPROM-Zellen der Ladungsverlust bei allen Temperaturen deutlich stärker, weil die untere Schwelle auch bei Temperaturen unterhalb von 400 °C im Beobachtungszeitraum fast im selben Maße zunimmt wie die obere Schwelle abnimmt. Dies liegt nicht am Substratmaterial sondern am dünneren Tunneloxid. Die neutrale Schwelle, die nach etwa 20 h bei 450 °C von den gelöschten und den programmierten Zellen gleichermaßen erreicht wird, befindet sich mit ungefähr 3,5 V nur wenig unterhalb der Mitte der anfänglichen gelöschten (7,5 V) und programmierten (0 V) Schwelle. Dennoch ist auch für die Bulk-EEPROM-Zellen der zeitliche Verlauf des Programmierfensters nicht logarithmisch und vor allem die Kurvenform nicht für alle Temperaturen identisch.

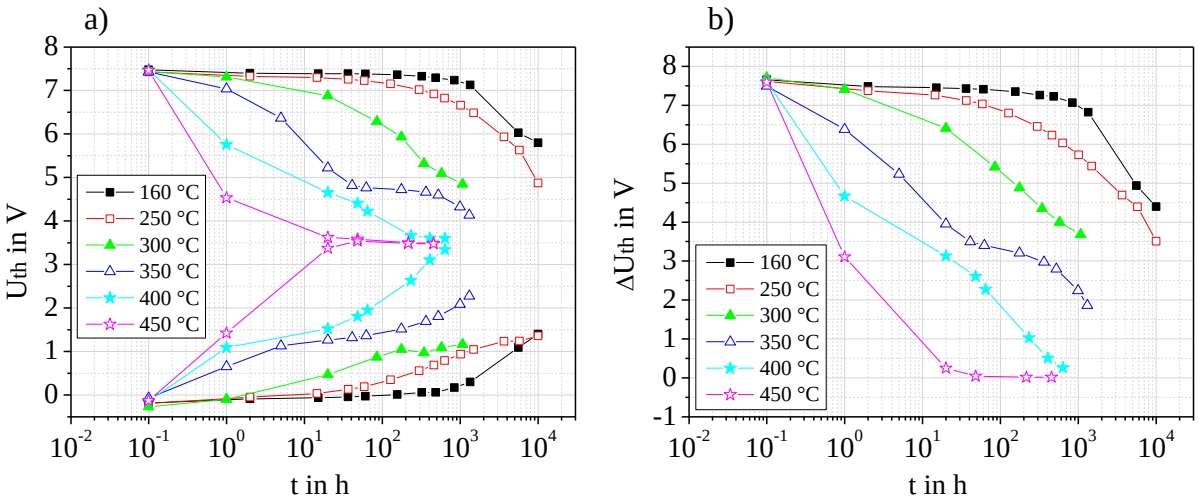


Abbildung 4.9: Schwellenspannungen für gelöschte (obere Schwellen) und programmierte (untere Schwellen) von auf Bulk-Wafern prozessierten EEPROM-Einzelzellen in Abhängigkeit von der Zeit bei Lagerungstemperaturen zwischen 160 °C und 450 °C (a); Differenz der oberen und der unteren Schwellen (Programmierfenster ΔU_{th}) in Abhängigkeit von der Zeit (b); die Messwerte bei t = 0 h sind wegen der logarithmischen x-Achsendarstellung bei t = 0,1 h eingezeichnet

Der zu Abbildung 4.9 gehörende Arrheniusgraph mit den Ausfallzeiten nach 15 % bzw. 25 % Änderung des Programmierfensters ist in Abbildung 4.10 dargestellt. Es fällt auf, dass die Ausfallzeiten nach 15 % Änderung des Programmierfensters deutlich weniger schwanken als bei den H10-EEPROM-Zellen, was daran liegt, dass sich alle sechs Kurven in Abbildung 4.9 sowohl bei der 15 % - als auch bei der 25 % - Änderung im stark abfallenden Kurvenabschnitt befinden. Wie auch in Abbildung 4.7 kann der Graph in zwei Bereiche mit einer niedrigeren und einer höheren Aktivierungsenergie aufgeteilt werden. Man erhält etwa 0,3 eV bzw. 1,3 eV in Bezug auf die 25 % - Änderung.

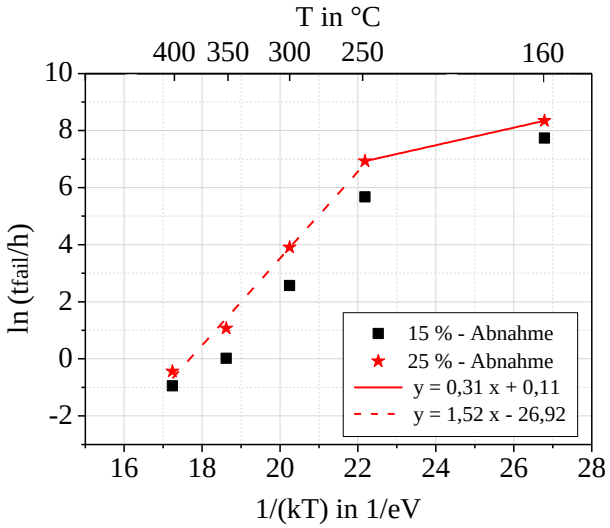


Abbildung 4.10: Logarithmus der Ausfallzeiten bei 15 % (Kästchen) bzw. 25 % (Sterne) Abnahme des Programmierfensters in Abhängigkeit von 1/(kT); die Messpunkte wurden aus Abbildung 4.9 b) gewonnen

Die Ergebnisse der Untersuchungen der Bulk-EEPROM-Zellen decken sich insofern mit den Beobachtungen der SOI-EEPROM-Zellen, als dass in beiden Fällen im Temperaturbereich

zwischen 250 °C und 450 °C die Aktivierungsenergie höher ist als üblicherweise angenommen[7], obwohl es sich in beiden Fällen um sehr unterschiedliche Zellen handelt (Substrat, Layout, Tunneloxiddicke, …). Weitere, hier nicht dargestellte Untersuchungen mit bei Raumtemperatur 10.000-mal vorgezykelten Zellen ergaben für beide EEPROM-Zellvarianten ebenso hohe Aktivierungsenergien[8]. Auch wenn sich die ermittelten Werte der Bulk-EEPROM-Zellen und der H10-EEPROM-Zellen unterscheiden und zudem von der betrachteten Änderung abhängen, geben sie doch insgesamt einen Hinweis darauf, dass der Ladungsverlust bei Temperaturen oberhalb von 250 °C auf einen Mechanismus zurückzuführen ist, der mit hohen Aktivierungsenergien in Zusammenhang steht. Dies wird im nächsten Abschnitt genauer untersucht.

Mobile Ionen im Backend: Vergleich der Prozessierung

In der Literatur findet sich eine große Bandbreite von Aktivierungsenergien im Zusammenhang mit dem Ladungserhalt von Speicherzellen. Niedrigere Aktivierungsenergien von 0,3 eV bis 0,6 eV werden für Oxid-Nitrid-Oxid-Strukturen [WU90], [PAN91], den Ausfall durch Oxiddefekte [SHI80] oder den dielektrischen Durchbruch der Tunneloxide [BAG84] angegeben. Für den Bereich mittlerer Aktivierungsenergien von 1 eV bis 1,2 eV werden durch Zyklen induzierte Defekte [VER88] oder Kontaminationen [SHI80] verantwortlich gemacht. Höhere Aktivierungsenergien von 1,4 eV bis 1,6 eV finden sich für den Fall des „intrinsischen Datenerhalts“, wo der Datenverlust nur eine Folge der Elektronenwanderung über die Potentialbarriere des Tunneloxids ist [HMP99], [HOU08]. Dies kommt aber nur dann zum Ausdruck, wenn die Oxide nicht vorgeschädigt oder defekt sind, weil sonst der intrinsische Datenverlust durch extrinsische Defekte (Oxiddefekte, Tunneloxiddurchbruch) überlagert werden würde.

Für noch höhere Aktivierungsenergien von 1,8 eV bis 2 eV, wie sie auch in dem hier vorliegenden Fall beobachtet werden konnten, kommt ein weiterer Mechanismus als Auslöser für den Datenverlust in Frage. Positive mobile Ionen, die während der Prozessierung des Back-Ends entstanden sind, können von einem negativ geladenen Floating-Gate angezogen werden. Als Folge sinkt die Schwellenspannung einer gelöschten EEPROM-Zelle, bei der ein Elektronenüberschuss auf dem Floating-Gate herrscht, während der Lagerung stark ab, während die programmierte Schwelle nahezu konstant bleibt [CRI90], [GAS00]. Dieser Mechanismus wird mit zunehmender Temperatur schnell stärker[9] und ist außerdem stark von der Prozessführung des Back-Ends abhängig.

Um zu überprüfen, ob der beschriebene Mechanismus auch für den hier vorliegenden Fall zutrifft, wurde der Datenerhalt von EEPROM-Zellen auf mehreren Wafern mit unterschiedlicher Back-End-Prozessierung verglichen. Die Ergebnisse sind in Abbildung 4.11 a) (gelöschte Schwellen) und b) (programmierte Schwellen) dargestellt. Die Wafer #2 bis #5 besitzen alle drei Lagen Wolframmetallisierung, unterscheiden sich aber in den Anlagen, die für die Prozessierung der Zwischenoxide und der Passivierung verwendet wurden. Die Prozessierung der Tunneloxide ist vergleichbar. Zum Vergleich wurden die Messwerte der

[7] Häufig wird von einer „Standardaktivierungsenergie“ für den Datenverlust von EEPROM-Zellen von 0,6 eV ausgegangen.

[8] Da keine Unterschiede für die Aktivierungsenergie von 10.000-mal vorgezykelten und „nur“ zehnmal gezykelten (initialisierten) Zellen vorlagen, wurden alle weiteren Experimente für diese Arbeit mit zehnmal vorgezykelten Zellen durchgeführt.

[9] Hohe Aktivierungsenergien bedeuten, dass mit zunehmender Temperatur die Ausfallzeit stärker sinkt als es bei einer niedrigeren Aktivierungsenergie der Fall wäre. Die Mechanismen, die zu sehr hohen Aktivierungsenergien führen, werden also stark durch Temperatur beschleunigt.

350 °C-Lagerung aus Abbildung 4.6 als Referenz (ausgefüllte schwarze Kästchen) eingezeichnet[10].

Da es nur um eine qualitative Gegenüberstellung geht, wurde die Lagerung bei nur einer Temperatur durchgeführt. Betrachtet man die Abbildungen 4.6 und 4.9, stellen sich dafür 350 °C als optimale Lagerungstemperatur heraus, bei der schon nach 24 h vorhandene Unterschiede im Datenerhalt deutlich werden können. Bei einer Lagerungstemperatur von 250 °C ist dies erst nach mehreren Hundert Stunden der Fall, bei Temperaturen von 400 °C und mehr ist der Ladungsverlust hingegen schon nach weniger als einer Stunde so groß, dass Unterschiede kaum noch auffallen.

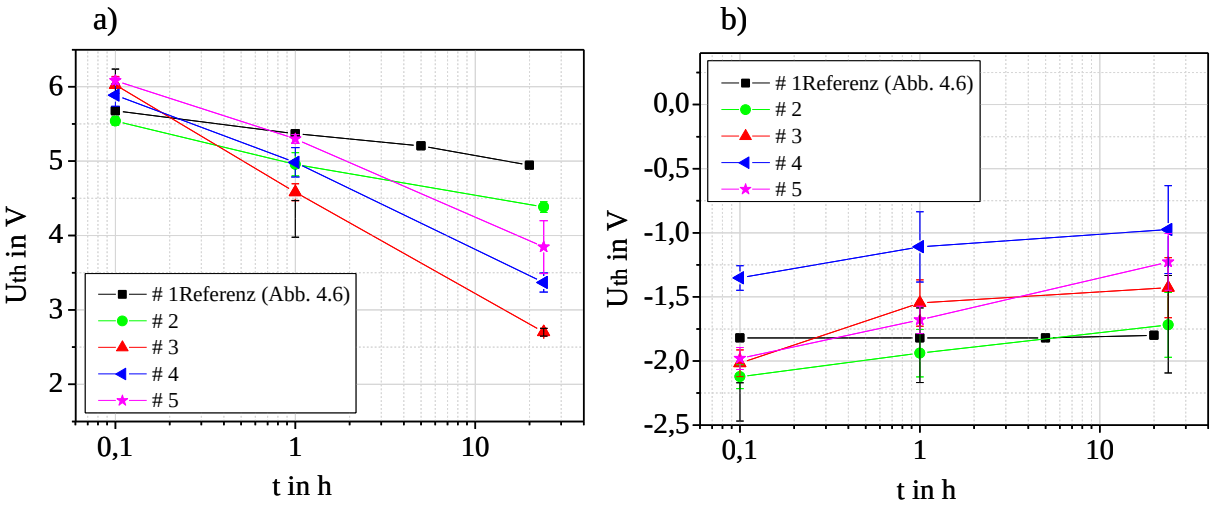


Abbildung 4.11: Schwellenspannungen für gelöschte (a) und programmierte Schwellen (b) von EEPROM-Einzelzellen in Abhängigkeit von der Zeit bei einer Lagerungstemperatur von 350 °C; Wafer mit unterschiedlicher Back-End-Prozessierung; die Messwerte des Referenzwafers #1 wurden aus Abbildung 4.6 a) entnommen; die Messwerte bei t = 0 h sind wegen der logarithmischen x-Achsendarstellung bei t = 0,1 h eingezeichnet; Mittelwerte aus bis zu zehn Einzelzellen mit Standardabweichung

In Abbildung 4.11 a) ist zu erkennen, dass trotz der anfänglich ungefähr auf demselben Niveau liegenden Schwellenspannungen die nach 24 h noch vorhandene Ladung von Wafer zu Wafer stark variiert. Die Anfangsschwellen der gelöschten Zellen unterscheiden sich leicht, wobei die Fälle starken Ladungsverlusts mit kleineren Anfangsschwellen korrelieren. Für die untere Schwelle in Abbildung 4.11 b) sind die Unterschiede nach 24 h deutlich geringer, für Wafer #4 liegt das Anfangsniveau aber etwa 0,5 V höher als in den anderen Fällen.

Die Prozessierung der Zwischenoxide und des Passivierungsoxids an verschiedenen Anlagen hat also einen entscheidenden Einfluss auf den Ladungserhalt. Wenn es sich tatsächlich um den Einfluss positiver Ionen handelt, erhöhen einige der Anlagen demnach die Zahl der positiven Ionen im Back-End, wodurch die obere Schwelle der EEPROM-Zellen mit der Zeit stärker sinkt als bei der Verwendung anderer Anlagen. Insgesamt ist es der Referenzwafer #1 aus Abbildung 4.6, bei dem an der gelöschten Schwelle der geringste Ladungsverlust auftritt und bei dem sich die untere Schwelle fast gar nicht ändert.

[10] Der Referenz-Wafer #1 hat eine zwei-lagige Aluminiummetallisierung. Die Metallisierung an sich und die Anzahl der Metalllagen haben aber keinen entscheidenden Einfluss auf den Datenerhalt, auch wenn für diesen Wafer der Ladungsverlust am geringsten ist.

Der Ladungsverlust wird also nicht nur vom Tunneloxid bestimmt, sondern auch vom Back-End. Dass es sich bei dem Grund für den starken Verlust tatsächlich um Ionen handelt, die positiv geladen und mobil sind, wie es bei Gassot *et al.* der Fall ist [GAS00], kann mit einem weiteren, von Gassot *et al*. vorgeschlagenen Experiment nachgewiesen werden.

Umprogrammierung der gelagerten Zellen

Dazu wurden neutrale EEPROM-Zellen auf den Wafern #2 und #3 aus Abbildung 4.11 initialisiert und die Hälfte der Zellen im gelöschten, die andere Hälfte im programmierten Zustand bei 350 °C gelagert[11]. Anschließend wurden die Zellen, die nun 24 h in gelöschtem Zustand lagerten, programmiert, und die Zellen, die 24 h in programmiertem Zustand waren, gelöscht und alle Zellen wieder 24 h gelagert. Abbildung 4.12 zeigt die Ergebnisse der zeitlichen Schwellenspannungsänderungen.

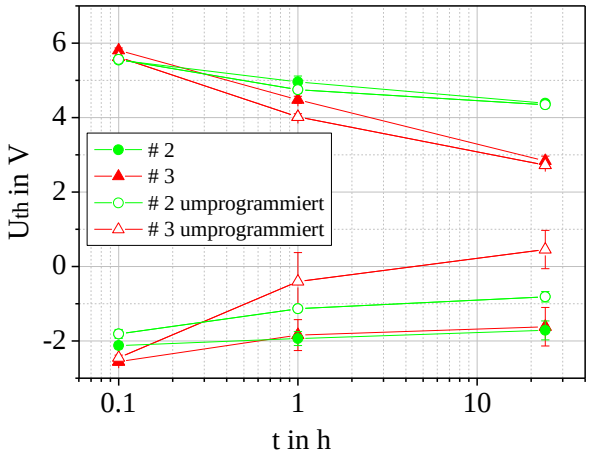


Abbildung 4.12: Schwellenspannungen für gelöschte und programmierte Schwellen von EEPROM-Einzelzellen in Abhängigkeit von der Zeit bei einer Lagerungstemperatur von 350 °C; erster Durchgang mit programmierten und gelöschten Zellen (ausgefüllte Symbole); zweiter Durchgang mit umprogrammierten Zellen (leere Symbole); auf jeweils zwei Wafern: #2 (Kreise) und #3 (Dreiecke); die Messwerte bei t = 0 h sind wegen der logarithmischen x-Achsendarstellung bei t = 0,1 h eingezeichnet; Mittelwerte aus je zehn Einzelzellen mit Standardabweichung

Während die obere Schwelle von Wafer #3 nach 24 h auf weniger als 3 V abgesunken ist liegt die entsprechende Schwelle von Wafer #2 etwa 1,5 V darüber, wie es auch schon in Abbildung 4.11 der Fall war. Die unteren Schwellen unterscheiden sich deutlich weniger. Nach der Umprogrammierung ergibt sich ein anderes Bild. Die Schwellenspannungen der vorher gelöschten und nun programmierten Zellen steigen deutlich stärker an als die anfangs programmierten Zellen es taten, vor allem die von Wafer #3. Je tiefer die obere Schwelle bei der ersten Lagerung sank, desto stärker steigt also die untere Schwelle bei der zweiten Lagerung an. Die oberen Schwellen der umprogrammierten EEPROM-Zellen verhalten sich aber ähnlich zu den anfangs gelöschten Zellen. Betrachtet man jeweils die obere und die untere Schwelle eines Wafers nach der Umprogrammierung, ist der zeitliche Verlauf beider Schwellen symmetrisch.

[11] Wafer #2 und Wafer #3 wurden ausgewählt, weil es sich dabei um den Wafer mit dem besten und den Wafer mit dem schlechtesten Datenerhalt in Abbildung 4.11 handelt.

Die Zunahme der unteren Schwelle nach der Umprogrammierung kann folgendermaßen erklärt werden: Wenn eine zuvor bei hoher Temperatur in gelöschtem Zustand gelagerte Zelle programmiert und dann erneut gelagert wird, steigt die Schwelle stark an, weil die Ionen, die bei der Temperaturlagerung der vorher gelöschten Zelle angezogen wurden, nun leicht abgestoßen werden können und dadurch die Ladung auf dem Floating-Gate schnell verändern [GAS00]. Dies bedeutet auch, dass die positiven Ionen, die vom negativ geladenen Floating-Gate angezogen wurden, nicht *auf* das Floating-Gate gelangen, sondern nur in dessen Nähe, sonst würden sie mit den Elektronen auf dem Floating-Gate rekombinieren und beim Umprogrammieren könnten sie nicht wieder abgestoßen werden. Eine zuvor im programmierten Zustand gelagerte Zelle, die dann gelöscht und erneut gelagert wird, verhält sich nicht anders als eine direkt gelöschte Zelle. Das bedeutet, dass negativ geladene Ionen im Back-End keinen Einfluss auf den Ladungserhalt haben.

Ursprung der mobilen Back-End-Ionen

Der Einfluss mobiler Ionen auf den Datenerhalt von Speicherzellen wurde von mehreren Gruppen diskutiert. Für den Ursprung der Ionen gibt es dabei unterschiedliche Interpretationen. Mielke *et al.* vermuten, dass Defekte zwischen zwei Polysiliziumlagen für den Datenverlust verantwortlich sind [MIE83]. Bei Crisenza *et al.* sind es mobile Ionen zwischen der zweiten Polyebene und der ersten Metallebene [CRI90] und bei Sakagami *et al.* wasserartige Gebilde zwischen den Metalllagen [SAK94], die jeweils einen negativen Einfluss auf den Datenerhalt haben. Die genannten Untersuchungen beziehen sich aber auf Doppel-Poly-Zellen. Gassot *et al.* arbeiteten hingegen mit Single-Poly-Zellen und führen die mobilen Ionen auf die Anwesenheit von Wasser im BPSG-Layer zurück [GAS00]. In den EEPROM-Zellen von Gassot *et al.* ist, wie auch in den in dieser Arbeit untersuchten Single-Poly-Zellen, das Floating-Gate direkt von der BPSG-Ebene umgeben, so dass positive Ionen im BPSG-Layer sehr einfach von einem negativ geladenen Floating-Gate angezogen werden könnten. Da bei uns aber nicht nur die Prozessierung des BPSG-Layers den Datenerhalt verändert sondern beispielweise auch die Prozessierung der Passivierung, könnten die mobilen Ionen ihren Ursprung auch in einer anderen Ebene haben.

Einen weiteren Hinweis auf die Anwesenheit zusätzlicher Ladungen (die aber keine Kontaminationen sind) geben die Ergebnisse von TVS-Messungen[12] am Gateoxid (Abbildung 4.13). Dabei wird an das Oxid mit Hilfe eines quasi-statischen Kapazitäts-Spannungs-Messgerätes eine Spannungsrampe von einer positiven zu einer negativen Spannung angelegt (jeweils obere Kurve), dann eine gewisse Zeit gewartet (etwa 90 s) und die Rampe zurück zur positiven Spannung gefahren (jeweils untere Kurve). Eigentlich dienen diese Messungen zur Detektierung von Kontaminationen (z. B. von Na^+- oder K^+-Ionen), hier ist aber ein anderer Effekt zu sehen[13]. Wafer #2 und #3 in Abbildung 4.13 entsprechen dabei Wafer #2 und #3 in Abbildung 4.11 und 4.12. Obwohl in dem niedrigen Potentialbereich keine Spannungsabhängigkeit des Stromes zu erwarten wäre, steigt der Strom mit betragsmäßig größer werdender Spannung an. Zusätzlich wird er nach dem Ende des ersten Teils der Spannungsrampe (bei U = -4 V) während der Haltezeit betragsmäßig noch größer. Die Leckströme sind größer, je höher die Temperatur ist und zudem für Wafer #3 (hoher Ladungsverlust) größer als für Wafer #2 (geringerer Ladungsverlust). Auch auf anderen Wafern, auf denen ein schlechter Ladungserhalt der EEPROM-Zellen nachgewiesen werden konnte, sind diese Leckströme besonders groß, d. h. die Fälle großer Leckströme in den TVS-Messungen

[12] TVS = Triangular Voltage Sweep (siehe JEDEC-Dokument [JED1])

[13] Sind Kontaminationen vorhanden, ist um U = 0 V ein Ionenpeak zu erkennen, was bei uns nicht der Fall ist. Die leichten Erhebungen im Stromverlauf sind nur eine Folge der Spannungsumpolung.

korrelieren mit den Fällen starken Ladungsverlustes in den Datenerhalt-Experimenten. In beiden Fällen ist die Prozessierung des Back-Ends der entscheidende Faktor.

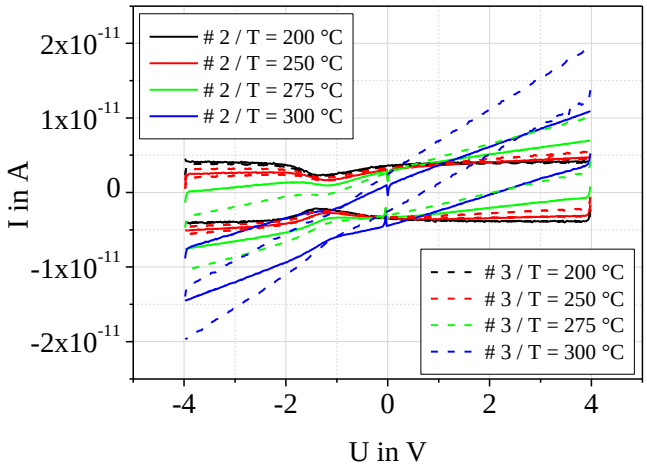


Abbildung 4.13: Quasi-statische IU-Kennlinien als TVS-Messungen an Gateoxidkondensatoren für verschiedenene Temperaturen (200 °C: schwarz, 250 °C: rot, 275 °C: grün, 300 °C: blau) auf zwei Wafern: #2 (durchgezogene Linien) und #3 (gestrichelte Linien); Wafer #2 entspricht Wafer #2 aus Abbildung 4.12 und Wafer #3 entspricht Wafer #3 aus Abbildung 4.12

Vergleich der Arrheniusgraphen und der Aktivierungsenergien für unterschiedlich prozessierte Wafer

Für die in Abbildung 4.11 gezeigten Beispiele müsste sich ein stärkerer Ladungsverlust in der Ausfallzeit, aber nicht in der Aktivierungsenergie selbst bemerkbar machen, wenn es sich um denselben physikalischen Mechanismus handelt, der nur in dem einen Fall stärker, in dem anderen Fall schwächer ist. Deshalb wurden weitere Lagerungen der beiden Wafer #2 und #3 bei drei verschiedenen Temperaturen vorgenommen[14]. Um eine größere Statistik zu erhalten, wurden statt Einzelzellen EEPROM-Arrays verwendet.

In Abbildung 4.14 a) sind die Änderungen der Schwellenspannungen in Abhängigkeit von der Zeit bei 250 °C, 350 °C und 400 °C dargestellt und in Abbildung 4.14 b) die jeweiligen zeitlichen Verläufe der Programmierfenster zu sehen. Abbildung 4.15 präsentiert die zugehörigen Arrheniusgraphen.

In Abbildung 4.14 a) kann man erkennen, dass die gelöschten EEPROM-Zellen von Wafer #3 nicht nur bei einer Lagerungstemperatur von 350 °C die Ladung schneller verlieren als die Zellen von Wafer #2, sondern auch bei 400 °C. Die gelöschten Schwellen bei einer Lagerungstemperatur von 250 °C sind für Wafer #3 hingegen zunächst größer als für Wafer #2, da auch die anfängliche Schwelle höher liegt. Betrachtet man Abbildung 4.14 b), so kann man erkennen, dass nach etwa 100 h die obere Schwelle von Wafer #3 stärker sinkt als die Schwelle von Wafer #2. Dieses Verhalten spiegelt sich auch in der Berechnung der Ausfallzeiten nach 15 % bzw. 25 % Änderung des Programmierfensters wider. Der Verlauf der Fit-Geraden für die 25 % - Abnahme ist für beide Wafer parallel, d. h. es ergeben sich die gleichen Aktivierungsenergien. Eine 15 % - Abnahme wird bei 250 °C in beiden Fällen bei

[14] Um keine Einfluss vorheriger Lagerungen zu haben, wurden die Wafer geviertelt und für jedes neue Experiment ein anderes Viertel verwendet.

ungefähr 2000 h erreicht, so dass sich die Fit-Geraden überschneiden. Hierbei wird wieder deutlich, wie entscheidend die Definition der Ausfallzeit die Ergebnisse beeinflusst.

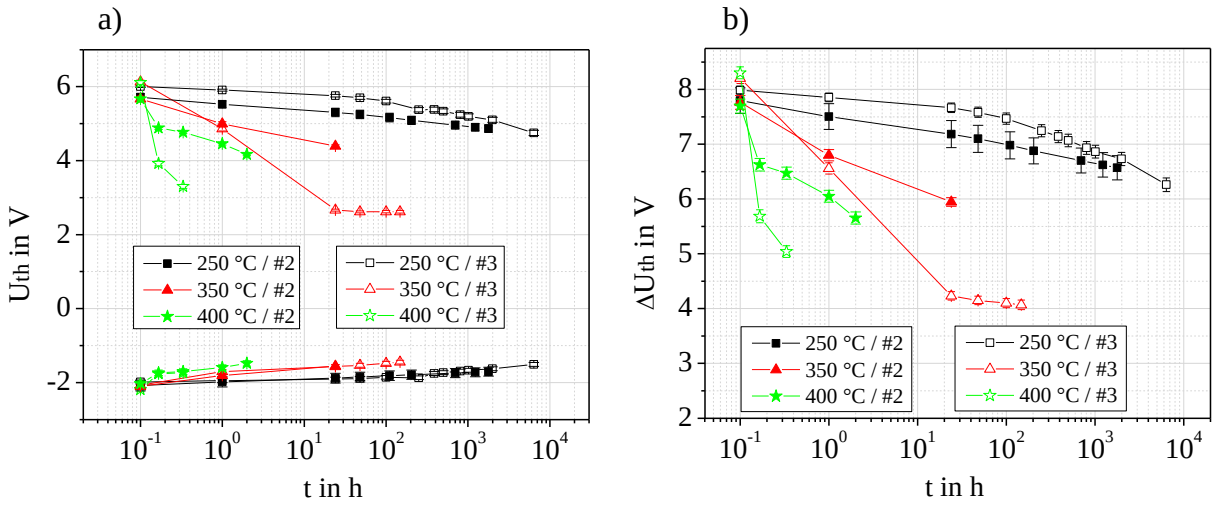


Abbildung 4.14: Schwellenspannungen für gelöschte (obere Schwellen) und programmierte (untere Schwellen) EEPROM-Einzelzellen in Abhängigkeit von der Zeit bei Lagerungstemperaturen zwischen 250 °C und 400 °C (a); Differenz der oberen und der unteren Schwellen (Programmierfenster ΔU_{th}) in Abhängigkeit von der Zeit (b); jeweils zwei Wafer: #2 (ausgefüllte Symbole) und #3 (leere Symbole); die Messwerte bei t = 0 h sind wegen der logarithmischen x-Achsendarstellung bei t = 0,1 h eingezeichnet worden

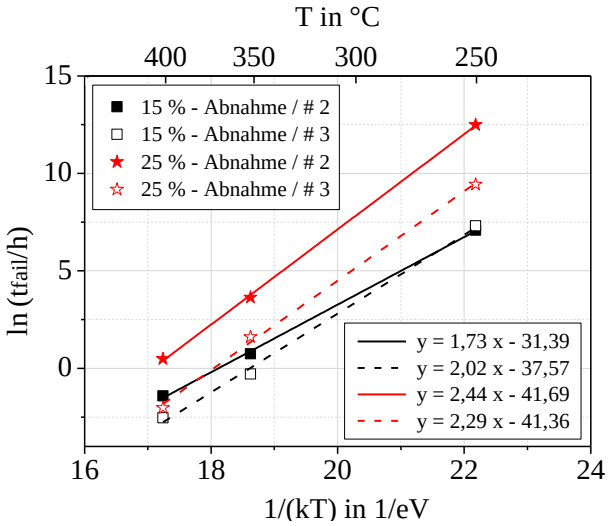


Abbildung 4.15: Logarithmus der Ausfallzeiten bei 15 % (Kästchen) bzw. 25 % (Sterne) Abnahme des Programmierfensters in Abhängigkeit von 1/(kT) für die beiden Wafer aus Abbildung 4.11: #2 (ausgefüllte Symbole) und #3 (leere Symbole); die Messpunkte wurden aus Abbildung 4.14 b) gewonnen

Mit etwa 1,7 bis 2,4 eV fügt sich die Aktivierungsenergie gut in die bisherigen Beobachtungen ein. Sie ist für einen Wafer mit relativ gutem Datenerhalt (#2) in etwa genauso groß wie für einen Wafer mit schlechterem Datenerhalt (#3), die Absolutwerte liegen für den schlechteren Wafer aber deutlich niedriger.

Der Verlauf der Kennlinien in Abbildung 4.14 zeigt auch, dass sich der „Ionen-Effekt“ bei 250 °C erst nach etwa 100 h auswirkt, wobei bis 2000 h die anfängliche Differenz der oberen

Schwellen der beiden Wafer immer noch einen größeren Einfluss hat als die mobilen Ionen. Dies bedeutet aber auch, dass der „Ionen-Effekt“ bei 250 °C erst nach einer Änderung des Programmierfensters von 15 % und mehr eine Rolle spielt. Bei einer Lagerungstemperatur von 160 °C würden sich die Ionen bei einer Aktivierungsenergie von etwa 2 eV erst nach einigen Jahren auswirken. Deshalb ist es schwierig, nachzuweisen, ob die ermittelte Aktivierungsenergie auch im Temperaturbereich unterhalb von 250 °C gilt.

Der Ionen-Effekt ist bei höherer Temperatur stärker, d. h. die Schwellenspannungsabnahme setzt bei höheren Temperaturen schon früher ein als bei niedrigeren Temperaturen. Ist eine bestimmte Ionenmenge in der Nähe des Floating-Gates vorhanden, tritt ein gewisser Abschirmungseffekt ein, d. h. das Anziehen weiterer Ionen wird erschwert. Dies erklärt auch, weshalb in Abbildung 4.6 bei 400 °C und 450 °C zunächst ein starker Abfall zu sehen ist, der dann schwächer wird. Zusätzlich ist natürlich auch bei einem kleinerem Programmierfenster die Potentialdifferenz zwischen Drain und Floating-Gate geringer, so dass weniger Elektronen tunneln können.

Fazit

- Zusammenfassend kann man nun festhalten, dass bei Temperaturen oberhalb von 250 °C positive mobile Ionen im Back-End für einen zusätzlichen Ladungsverlust der EEPROM-Zellen, vor allem ihrer gelöschten Schwellen, veranwortlich sind. Hinweise auf einen Zusammenhang mit dem Back-End geben Vergleiche verschiedener Wafer. Eine größere Ionenzahl, d. h. ein stärkerer Ladungsverlust, hat kleinere Ausfallzeiten zur Folge, die Aktivierungsenergie ändert sich aber nicht.

- Bei Temperaturen oberhalb von 250 °C ergeben sich in einem Arrheniusgraphen Aktivierungsenergien von etwa 2 eV. Die genaue Bestimmung ist schwierig, da bei den verschiedenen Temperaturen die Schwellenspannungsänderungen unterschiedlich verlaufen. Dadurch kann die berechnete Aktivierungsenergie für verschiedene prozentuale Abnahmen des Programmierfensters zur Bestimmung der Ausfallzeit auch unterschiedlich sein. Für Temperaturen unterhalb von 250 °C lassen sich die Ionen nicht nachweisen, da hier die Ausfallzeiten sehr groß sind.

- Für schnelle Vergleiche von EEPROM-Zellen auf verschiedenen Wafern ist eine Lagerungstemperatur von 350 °C gut geeignet, weil schon innerhalb von 24 h Unterschiede sehr deutlich werden.

- Die Lagerung der EEPROM-Zellen beeinflusst die auf dem Floating-Gate gespeicherte Ladung, zerstört die Zellen an sich aber nicht. Auch eine mehrere Stunden bei 450 °C gelagerte Zelle kann weiterhin programmiert und gelöscht werden. Eine erneute Lagerung wird aber von den vorherigen Lagerungsexperimenten beeinflusst, vor allem, wenn die Zelle vorher im gelöschten Zustand war. Deshalb ist es wichtig, dass Messungen zum Datenerhalt nur an Wafern vorgenommen werden, die vorher noch keine Temperaturlagerung erfahren haben, weil nicht auszuschließen ist, dass Zellen schon mit einer gewissen Ladungsmenge auf dem Floating-Gate aus der Fertigung kommen.

- Bei einer Betriebstemperatur von 250 °C liegt die Lebensdauer der EEPROM-Zellen für eine 25 % - Abnahme des ursprünglichen Programmierfensters zwischen 10.000 h und 200.000 h, je nachdem, wie die Back-End-Prozessführung und damit der Einfluss der Ionen ist.

4.3.2 Zyklenfestigkeit (Endurance)

Der zweite Zuverlässigkeitsaspekt, die Zyklenfestigkeit oder Endurance, spielt im Hinblick auf die Nutzung der EEPROM-Zellen eine ebenso wichtige Rolle wie der Datenerhalt. Im Folgenden wird vor allem die Abhängigkeit einer als maximal möglich definierten Zyklenzahl von der Temperatur, aber auch von der Programmierspannung und der Tunneloxiddicke untersucht. Die Messungen wurden an Einzelzellen bei Temperaturen zwischen -40 °C und 250 °C auf Waferebene und bei Temperaturen zwischen 300 °C und 400 °C im Gehäuse durchgeführt.

Endurance-Kurven in Abhängigkeit von der Programmierspannung

In Abbildung 4.16 sind Endurance-Kurven, d. h. die jeweilige gelöschte und programmierte Schwelle in Abhängigkeit von der Anzahl der Zyklen für Programmierspannungen U_{pp} zwischen 13 V und 19 V bei 25 °C (a) und 250 °C (b) dargestellt. Alle EEPROM-Zellen wurden, sofern die Schwellen lesbar waren, mindestens eine Million Mal gezykelt.

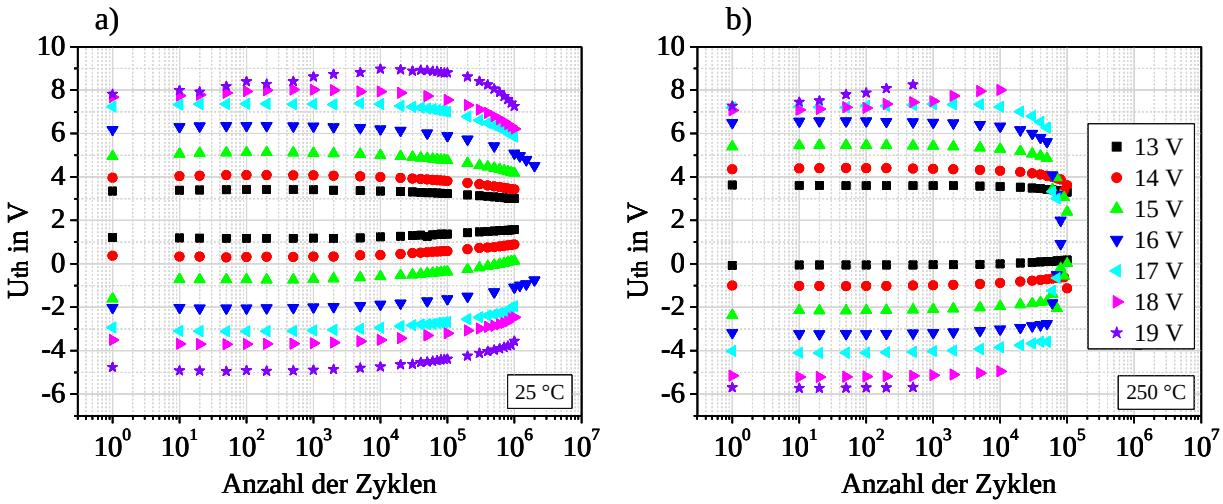


Abbildung 4.16: Schwellenspannungen in Abhängigkeit von der Anzahl der Zyklen für Programmierspannungen U_{pp} zwischen 13 V und 19 V bei 25 °C (a) und 250 °C (b); Tunneloxiddicke d_{tox} = 10,9 nm; die Legende in Abbildung 4.16 (b) bezieht sich auch auf Abbildung 4.16 (a)

Je größer die Programmierspannung ist, desto größer ist auch das Programmierfenster zu Beginn der Messung, was vor allem bei der unteren Schwelle deutlich wird. Je größer die Programmierspannung ist, desto schneller nähern sich aber auch die Schwellen mit zunehmender Zyklenzahl an. Bei 250 °C ist dies daran zu erkennen, dass bei einer größeren Programmierspannung weniger Zyklen möglich sind, bevor die Zelle zerstört wird (z. B. für U_{pp} = 16 V nach ca. 80.000 Zyklen, während für U_{pp} = 14 V die Zelle auch nach 100.000 Zyklen noch intakt ist) oder die Schwellen einfach gar nicht mehr lesbar sind (U_{pp} = 18 V nach ca. 10.000 Zyklen und U_{pp} = 19 V nach ca. 500 Zyklen). Dafür ist die Aufweitung des Programmierfensters aufgrund des positiven Ladungseinfangs (siehe Abschnitt 4.1.4) bei den größeren Programmierspannungen auch stärker ausgeprägt. Besonders an der oberen Schwelle ist dies zu erkennen.

Eine größere Programmierspannung bedeutet, dass das über dem Oxid anliegende elektrische Feld auch größer ist. Dadurch ist es möglich, beim Programmieren bzw. Löschen mehr Ladung vom Floating-Gate herunter zu ziehen bzw. auf dieses zu bringen. Folglich weitet sich das Programmierfenster. Dieser Effekt ist stärker, wenn Elektronen vom Floating-Gate herunter gezogen werden (Entstehen der unteren Schwelle), d. h. wenn Elektronen vom Polysilizium zum Aktivgebiet (Silizium) tunneln statt umgekehrt (vergleiche Abbildung 4.4). Für Elektronen ist es einfacher, die Potentialbarriere auf Seite des Polysiliziums zu überwinden als vom Silizium zum Polysilizium zu tunneln, da die Dotierung der Siliziumgebiete geringer ist als die des Polysiliziums, wo die $POCl_3$-Belegung (Phosphoroxychlorid) zu einem deutlich höheren Wert führt (siehe auch Kapitel 5).

Ein größeres elektrisches Feld weitet aber nicht nur das Programmierfenster, sondern setzt das Tunneloxid auch einem größeren Stress aus, wodurch es schneller degradiert und die Zelle schneller zerstört wird.

Betrachtet man das Programmierfenster nach einem Zyklus genauer, fällt auf, dass die obere Schwelle bei U_{pp} = 19 V nur geringfügig über der Schwelle von U_{pp} = 17 V und U_{pp} = 18 V liegt. Diese Beobachtung stimmt mit dem Sättigungseffekt überein, der in Abbildung 4.5 zu sehen ist. Nach mehrfachem Zykeln aber nimmt die obere Schwelle bei den größeren Programmierspannungen deutlich zu, bevor sie letztlich absinkt, d. h. die Sättigung kann durch mehrfaches Programmieren und Löschen aufgehoben werden.

Endurance-Kurven in Abhängigkeit von der Temperatur

In Abbildung 4.16 ist zu erkennen, dass bei der gleichen Programmierspannung die unteren Schwellen bei 250 °C tiefer liegen als bei 25 °C, während die oberen Schwellen nahezu gleich sind. Um die Abhängigkeiten von der Temperatur genauer zu betrachten, sind in Abbildung 4.17 vier Diagramme dargestellt, die jeweils Endurance-Kurven bei Temperaturen zwischen -40 °C und 250 °C beinhalten. Die Tunneloxiddicke der EEPROM-Zellen in den Abbildungen a) und b) beträgt 10,9 nm und in den Abbildungen c) und d) 13,2 nm. Die Zellen in den Abbildungen a) und c) wurden mit U_{pp} = 16 V, die Zellen in den Abbildungen b) und d) bei U_{pp} = 19 V gezykelt.

Wie auch schon in Abbildung 4.16 zu sehen war, ist die untere Schwelle nach einem Zyklus bei einer höheren Temperatur niedriger als bei einer tieferen Temperatur, wobei der Einfluss der Temperatur auf das Programmierfenster deutlich geringer ist als der Einfluss der Programmierspannung oder der Tunneloxiddicke. Vor allem aber läuft das Programmierfenster bei einer höheren Temperatur schneller und abrupter zusammen als bei einer niedrigeren Temperatur, d. h. die Zelle wird mit zunehmender Temperatur schneller zerstört.

Die Beobachtungen, die in Abbildung 4.16 bezüglich der Abhängigkeit des Programmierfensters von der Programmierspannung gemacht werden konnten, treffen auch auf die Abhängigkeit von der Tunneloxiddicke zu, da in beiden Fällen das über dem Oxid anliegende elektrische Feld beeinflusst wird. Für alle Temperaturen ist das Programmierfenster zu Beginn der Messung bei einem dünneren Tunneloxid größer als bei einem dickeren Tunneloxid. Dafür weitet sich das Programmierfenster bei dem dickeren Tunneloxid aber weniger auf und sinkt später weniger steil ab.

In den meisten Endurance-Kurven unterscheiden sich die Schwellenspannungsniveaus nach einem Zyklus und nach zehn Zyklen kaum. Bei der Betrachtung von Abbildung 4.17 c) fällt

aber auf, dass bei der 100 °C - Kurve (nach oben zeigende Dreiecke) der Messwert der unteren Schwelle nach einem Zyklus etwa 1 V tiefer liegt als der Wert nach zehn Zyklen. Der Grund hierfür liegt darin, dass auf dem Floating-Gate dieser EEPROM-Zelle schon nach ihrer Herstellung ein Mangel an Elektronen vorlag, d. h. die Zelle schon im programmierten Zustand war, bevor sie erneut programmiert wurde[15]. Dadurch konnte sich das Schwellenspannungsniveau weiter herabsenken. Nachdem die Zelle aber mehrfach gelöscht und wieder programmiert wurde, hat sich das Niveau der Schwellen den Zellen angepasst, die nach ihrer Herstellung im neutralen Zustand waren. Daraus ergibt sich die Konsequenz, dass für eine stabile Anfangsschwelle, wie sie beispielsweise bei den Experimenten zum Datenerhalt nötig ist, ein zehnmaliges Zykeln (Initialisieren) sinnvoll ist.

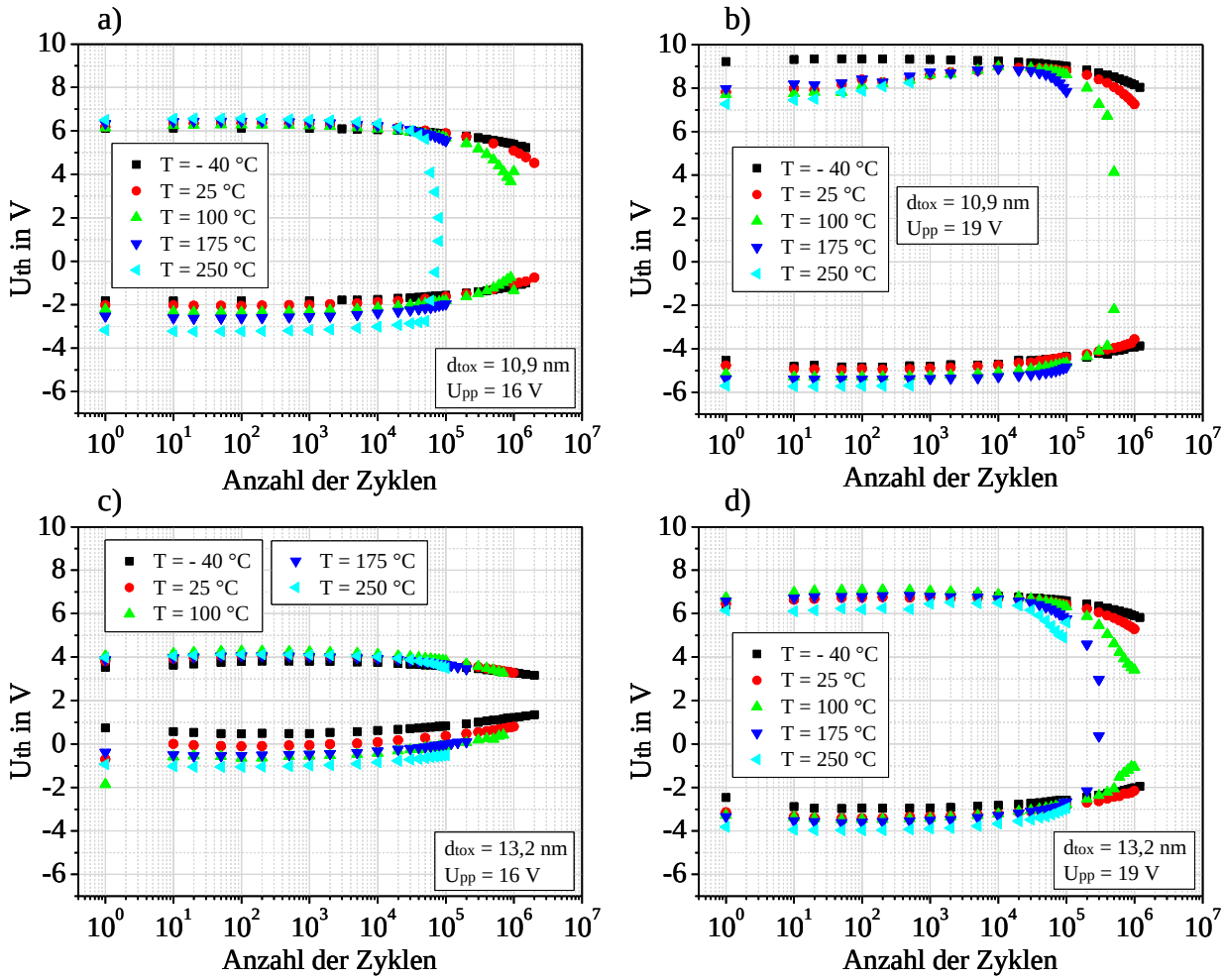


Abbildung 4.17: Schwellenspannungen in Abhängigkeit von der Anzahl der Zyklen für Temperaturen zwischen -40 °C und 250 °C; Programmierspannungen U_{pp} = 16 V (a und c) bzw. U_{pp} = 19 V (b und d); Tunneloxiddicken d_{tox} = 10,9 nm (a und b) bzw. d_{tox} = 13,2 nm (c und d)

Für das mit zunehmender Temperatur leicht größere Programmierfenster zu Beginn der Messung ist der Anstieg des Fowler-Nordheim-Stromes mit der Temperatur verantwortlich. Je größer die Temperatur ist, desto leichter ist es, Elektronen auf das Floating-Gate zu bringen bzw. sie herunter zu ziehen. Die Temperaturabhängigkeit des Fowler-Nordheim-Stromes ist zudem für kleinere Felder größer [SUN93], so dass dieser Effekt bei U_{pp} = 16 V stärker ist als

[15] Bei einem Zyklus wird die EEPROM-Zelle immer erst programmiert und dann gelöscht.

bei U_{pp} = 19 V. Mit der Temperatur degradiert die Zelle aber auch schneller, da das Tunneloxid bei einer höheren Temperatur einem größeren Stress ausgesetzt ist, so dass die Schwellen mit zunehmender Zyklenzahl abrupter zusammenlaufen.

Dass die Erhöhung der Temperatur wie auch die Erhöhung der Programmierspannung einen stärkeren Einfluss auf das Anfangsniveau der unteren als der oberen Schwelle hat, kann wieder mit dem größeren Elektronenfluss vom Polysilizium zum Silizium erklärt werden (siehe oben).

Zur Ergänzung von Abbildung 4.17 sind in Abbildung 4.18 Endurance-Kurven bei Temperaturen bis 400 °C dargestellt. Die Programmierspannung bei den Messungen betrug 16 V und das Tunneloxid war 11,7 nm dick.

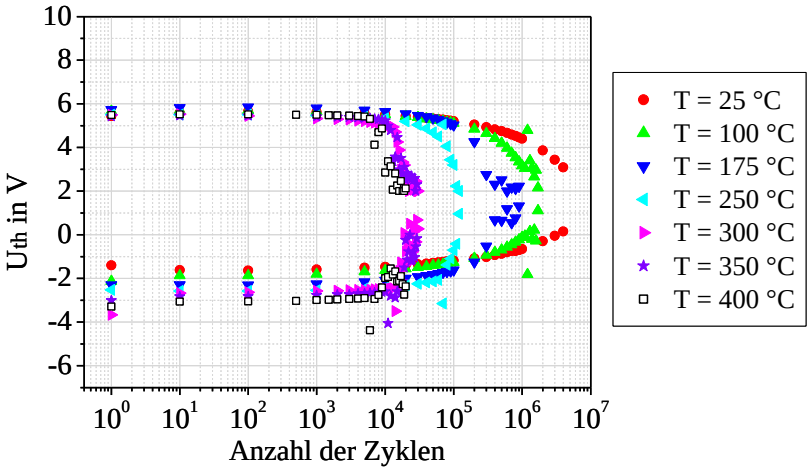


Abbildung 4.18: Schwellenspannungen in Abhängigkeit von der Anzahl der Zyklen für Temperaturen zwischen 25 °C und 400 °C; Programmierspannung U_{pp} = 16 V; Tunneloxiddicke d_{tox} = 11,7 nm

Der Temperatureinfluss auf das Anfangsfenster, der zwischen -40 °C und 250 °C beobachtet werden konnte (Abbildung 4.17), setzt sich bis 400 °C fort. Deutlicher als in den bisherigen Abbildungen ist hier zudem das frühere Zusammenlaufen der oberen und der unteren Schwellen mit größer werdender Temperatur zu erkennen. Während sich zwischen 25 °C und 250 °C die Zyklenzahl, bei der die gelöschte und die programmierte Schwelle nicht mehr unterscheidbar sind, alle 75 °C um fast eine Dekade reduziert, kann dies ab 350 °C nicht mehr so genau beobachtet werden. Dies liegt daran, dass bei diesen hohen Temperaturen die Schwellen nicht gleichmäßig zusammenlaufen, sondern ab einer bestimmten Temperatur die Schwellenspannungswerte stark schwanken.

Abbildung 4.19 zeigt für 25 °C, 250 °C und 400 °C die Abhängigkeit des Drain-Source-Stromes von der angelegten Control-Gate-Spannung beim Lesen der gelöschten Schwellen. Zwar ist bei 400 °C das Leckstromniveau mit einigen 10 nA sehr hoch, die Schwelle bei 1 µA kann aber problemlos gelesen werden. Für Temperaturen oberhalb von 400 °C wird die Bestimmung der Schwellenspannung aber zunehmend schwierig (vergleiche dazu auch Kapitel 6.1.2).

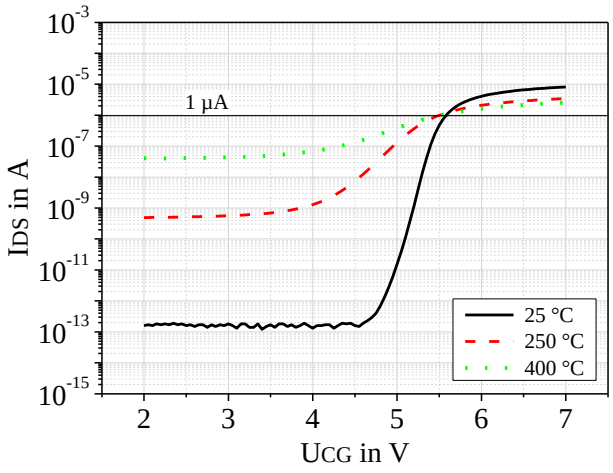


Abbildung 4.19: Drain-Source-Strom nach 10 Zyklen in Abhängigkeit von der Control-Gate-Spannung beim Lesen der gelöschten Schwellen bei 25 °C (durchgezogene Linie), 250 °C (gestrichelte Linie) und 400 °C (gepunktete Linie)

Zusammenfassend lässt sich festhalten, dass eine höhere Temperatur die Zyklenzahl reduziert, bei der die obere und die untere Schwelle zusammen laufen, obwohl das anfängliche Programmierfenster größer ist. Bezüglich der Abhängigkeit von der Programmierspannung und der Tunneloxiddicke kann man aus den Abbildungen entnehmen, dass eine größere Programmierspannung oder ein dünneres Tunneloxid das anfängliche Programmierfenster stärker vergrößern und die Schwellen früher und abrupter zusammenlaufen lassen als es eine kleinere Programmierspannung oder ein dickeres Tunneloxid bewirken können.

Definition des Ausfallkriteriums: prozentuale versus absolute Abnahme

Um die Abhängigkeit der Entwicklung des Programmierfensters mit zunehmender Zyklenzahl von der Programmierspannung, der Tunneloxiddicke oder der Temperatur quantitativ zu betrachten, muss ähnlich wie bei der Bestimmung einer Ausfallzeit eine maximale Zyklenzahl definiert werden, bei der das Kriterium, das eine zuverlässige Zelle kennzeichnet, nicht mehr erfüllt ist. Dies müsste eigentlich das minimale Programmierfenster sein, bei dem man die gelöschte Schwelle noch sicher von der programmierten Schwelle unterscheiden kann. Da nicht in allen Messungen so lange gezykelt werden konnte, bis die Schwellen nicht mehr voneinander zu unterscheiden waren[16], ist es nicht möglich, ein minimales Programmierfenster als Ausfallkriterium festzulegen. Im Folgenden werden deshalb in Analogie zur Definition des Ausfallkriteriums beim Datenerhalt zwei andere Möglichkeiten betrachtet: Die Definition des Ausfallkriteriums als prozentuale oder als absolute Änderung des anfänglichen Programmierfensters[17]. Wird im Folgenden die Bezeichnung „maximale Zyklenzahl“ verwendet, so handelt es sich um die Zyklenzahl, bei der eine bestimmte (festgelegte) Degradation erreicht wird, aber nicht zwangsläufig um die Zyklenzahl, nach der die Zelle definitiv defekt ist.

Im Gegensatz zu den Untersuchungen zum Datenerhalt (Abschnitt 4.3.1) ist bei der Zyklenfestigkeit die Tatsache zu berücksichtigen, dass das Programmierfenster zu Beginn der

[16] Vor allem bei den niedrigeren Temperaturen wäre dies nämlich erst nach einigen Millionen Zyklen der Fall.
[17] Wegen der teilweise auftretenden Instabilität der Schwellenspannungen nach nur einem Zyklus wird das anfängliche Fenster im Folgenden immer als das Fenster nach zehn Zyklen interpretiert.

Messung je nach Temperatur, Tunneloxiddicke und Programmierspannung unterschiedlich groß ist. Eine prozentuale Abnahme hat also einen anderen Effekt als eine absolute Abnahme. Diese Diskrepanz wird deutlich, wenn man die Entwicklung des Programmierfensters mit zunehmender Zyklenzahl bei verschiedenen Bedingungen genauer betrachtet. In Abbildung 4.20 sind dazu Kombinationen von zwei Temperaturen und zwei Programmierspannungen dargestellt. Die Auswahl ist auch repräsentativ für andere Spannungen und Tunneloxiddicken.

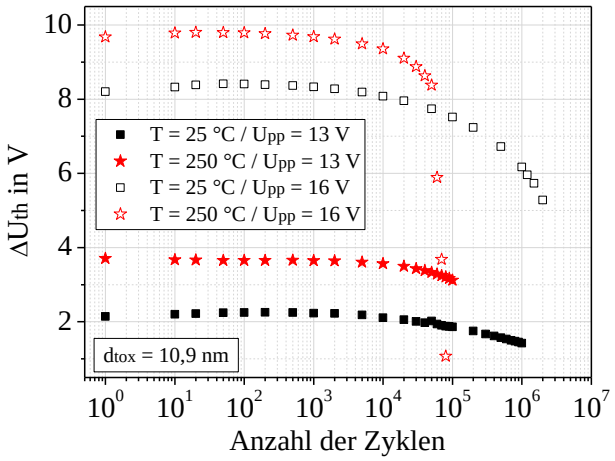


Abbildung 4.20: Programmierfenster in Abhängigkeit von der Anzahl der Zyklen für Programmierspannungen U_{pp} von 13 V (ausgefüllte Symbole) und 16 V (leere Symbole) bei 25 °C (Kästchen) und 250 °C (Sterne); Tunneloxiddicke d_{tox} = 10,9 nm

Zur genaueren Analyse der Zyklenzahlen, nach denen eine bestimmte prozentuale oder absolute Abnahme des anfänglichen Programmierfensters erreicht wird, sind in Tabelle 4.2 einige Informationen, die aus Abbildung 4.20 gewonnen wurden, zusammengestellt. Die prozentuale Abnahme von 15 % und die absolute Abnahme von 1 V wurden gewählt, weil dies Abnahmen sind, die in (fast) allen Fällen erreicht wurden.

***T* in °C**	**25**	**250**	**25**	**250**
U_{pp} in V	**13**	**13**	**16**	**16**
ΔU_{th} in V	2,1	3,7	8,2	9,7
-15 %	ca. 1,8	ca. 3,1	ca. 7,0	ca. 8,2
***Z* 15%**	***ca. 150.000***	***ca. 100.000***	***ca. 300.000***	***ca. 50.000***
-1 V	1,1	2,7	7,2	8,7
***Z* 1V**	***> 1 Mio.***	***> 100.000***	***ca. 200.000***	***ca. 40.000***

Tabelle 4.2: Gegenüberstellung prozentualer und absoluter Abnahmen des anfänglichen Programmierfensters für die vier Temperatur-Programmierspannungs-Kombinationen aus Abbildung 4.20; Z bezeichnet die Zyklenzahl

Vergleicht man die möglichen Zyklenzahlen bei 25 °C, so fällt auf, dass bei Betrachtung der prozentualen Abnahme bei U_{pp} = 13 V nur etwa die Hälfte der Zyklen erreicht werden, die bei U_{pp} = 16 V möglich sind. Bei der größeren Programmierspannung sind also mehr Zyklen

möglich, was allein am größeren Programmierfenster und der dadurch größeren möglichen Spannungsabnahme liegt. Für die absolute Abnahme aber liegt die Zyklenzahl mit mehr als 1 Mio. bei U_{pp} = 13 V deutlich über dem Ergebnis bei U_{pp} = 16 V. Die prozentuale und die absolute Abnahme liefern also ein anderes Resultat bezüglich der Abhängigkeit der Zyklenzahl von der Programmierspannung.

Die Gegenüberstellung der beiden Programmierspannungen bei 250 °C unterscheidet sich von den Beobachtungen bei 25 °C dahingehend, dass sowohl bei Betrachtung der prozentualen als auch der absoluten Abnahme bei U_{pp} = 13 V deutlich mehr Zyklen möglich sind als bei U_{pp} = 16 V. Ausschlaggebend ist hier nicht die Größe des Programmierfensters, sondern der Kurvenverlauf, der bei U_{pp} = 16 V deutlich steiler ist als bei U_{pp} = 13 V.

Die Wahl des Auswahlkriteriums ist also entscheidend, bei welcher Bedingung wie viele Zyklen möglich sind. Die Definition des Ausfallkriteriums muss eigentlich an die Anforderungen des Schaltungsdesigns zur Trennung gelöschter und programmierter Zellen angepasst werden. Dies betrifft aber nur einen kleineren Programmierspannungsbereich (15 V bis 17 V). Da in dieser Arbeit aber der Anspruch besteht, einen größeren Spannungs- und Temperaturbereich zu untersuchen, werden im Folgenden die Abhängigkeit der Zyklenzahl von der Programmierspannung, der Tunneloxiddicke und der Temperatur bei einer prozentualen und einer absoluten Abnahme quantitativ betrachtet. Der verwendete Begriff „maximale Zyklenzahl“ ist also nicht als das schaltungstechnisch Mögliche, sondern in Bezug auf die betrachtete Degradation zu verstehen.

Abhängigkeit der „maximalen Zyklenzahl“ von der Programmierspannung und der Tunneloxiddicke

Abbildung 4.21 zeigt den Logarithmus der „maximalen Zyklenzahl“ in Abhängigkeit von der Programmierspannung bei 25 °C und 250 °C. Dabei ist die „maximale Zyklenzahl“ als eine Abnahme von 15 % (a) bzw. 1 V (b) des anfänglichen Programmierfensters definiert. Die den Datenpunkten zu Grunde liegenden Messungen sind in Abbildung 4.16 zu sehen. Bei 250 °C konnten für U_{pp} > 17 V keine Werte eingezeichnet werden, da die Endurance-Kurven abrupt abbrechen, bevor eine Änderung von 15 % bzw. 1 V erreicht wurde (siehe Abbildung 4.16).

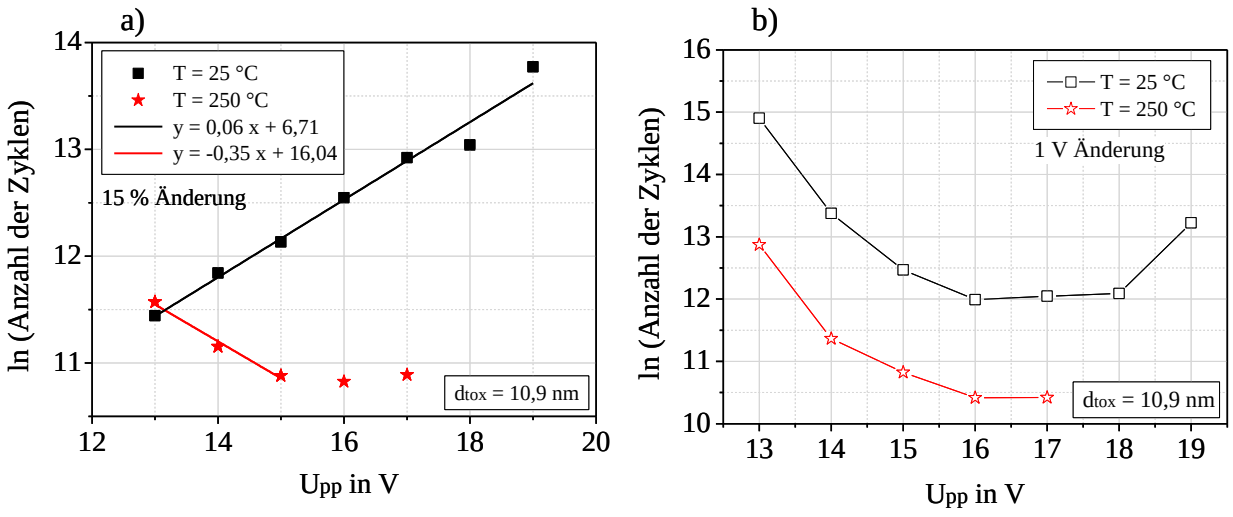


Abbildung 4.21: Logarithmus der Anzahl der Zyklen nach 15 % Änderung (a) bzw. 1 V Änderung (b) des Programmierfensters in Abhängigkeit von der Programmierspannung bei konstanter Tunneloxiddicke d_{tox} = 10,9 nm; T = 25 °C (Kästchen) und T = 250 °C (Sterne)

Der Kurvenverlauf für die absoluten Änderungen ist bei 25 °C und 250 °C parallel, die Zyklenzahlen sind aber bei 250 °C kleiner. Für größere Programmierspannungen sind weniger Zyklen möglich, da das Fenster schneller kleiner wird als bei kleineren Programmierspannungen, wobei sich ab U_{pp} = 16 V bei beiden Temperaturen die Zyklenzahlen kaum noch unterscheiden (vergleiche dazu Abbildung 4.16 b). Eine Ausnahme bildet die Programmierspannung von 19 V bei 25 °C. Hier sind deutlich mehr Zyklen möglich als bei U_{pp} = 18 V, weil sich die Abnahme von 1 V genau in dem Bereich befindet, in dem die steile Abnahme noch nicht eingesetzt hat, die anfängliche Aufweitung aber zunächst für eine Zunahme des Fensters gesorgt hat, bevor es um 1 V abgenommen hat. Würde man statt 1 V eine Abnahme von 2 V betrachten, wären bei U_{pp} = 19 V weniger Zyklen möglich als bei U_{pp} = 18 V. Da aber nicht in allen Fällen eine Abnahme von 2 V erreicht wurde, sind diese Werte nicht dargestellt.

Bei Betrachtung der prozentualen Abnahme liegt ein unterschiedliches Verhalten für 25 °C und 250 °C vor. Während bei 250 °C die Zyklenzahl mit steigender Programmierspannung zunächst abnimmt und dann relativ konstant bleibt, wird sie bei 25 °C mit steigendem U_{pp} immer größer. Dieses Verhalten konnte auch schon in Abbildung 4.20 beobachtet werden und liegt daran, dass bei einer größeren Programmierspannung auch das Programmierfenster größer ist und dadurch eine größere Spannungsabnahme erlaubt ist. Interessant ist aber vor allem die Tatsache, dass der Verlauf im gesamten betrachteten Bereich exponentiell ist. Dies bedeutet, dass man einen Zusammenhang zwischen der Anzahl der Zyklen $Z_{15\%}$ nach 15 % Änderung des Programmierfensters und der Programmierspannung aufstellen kann:

$$\ln(Z_{15\%}) = a \cdot U_{pp} + b \qquad \text{bzw.} \qquad Z_{15\%} = b' \cdot \exp(a \cdot U_{pp}) \qquad [4.3]$$

Die Konstanten *a*, *b* und *b*’ sind temperaturabhängig. Für 25 °C ist a positiv und für 250 °C negativ. Um genauer zu analysieren, wie sich dieser Zusammenhang für Temperaturen zwischen 25 °C und 250 °C verhält, ist in Abbildung 4.22 der Logarithmus der „maximalen Zyklenzahl“ bei 15 % Abnahme des Programmierfensters in Abhängigkeit von der Tunneloxiddicke aufgetragen.

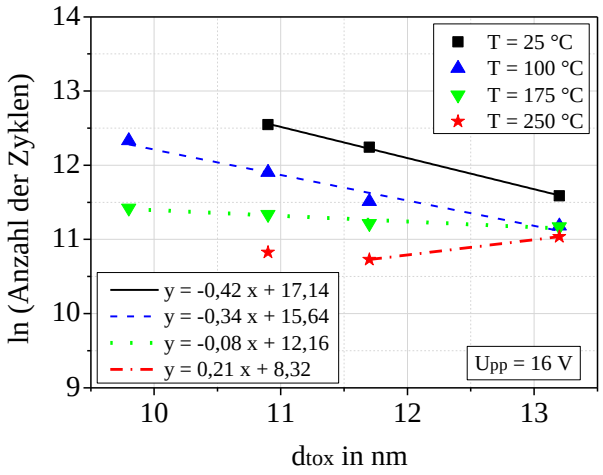


Abbildung 4.22: Logarithmus der Anzahl der Zyklen nach 15 % Änderung des Programmierfensters in Abhängigkeit von der Tunneloxiddicke bei konstanter Programmierspannung U_{pp} = 16 V; *T* = 25 °C (Kästchen), *T* = 100 °C (nach oben zeigende Dreiecke); *T* = 175 °C (nach unten zeigende Dreiecke) und *T* = 250 °C (Sterne)

Es zeigt sich, dass die Zusammenhänge aus Gleichung 4.3 auch für die Abhängigkeit der Zyklenzahl von der Tunneloxiddicke gelten und dabei der Übergang der Vorzeichen für die Steigung bei etwa 175 °C liegt. Die Konstanten *c*, *d* und *d*' sind auch hier temperaturabhängig.

$$\ln(Z_{15\%}) = c \cdot d_{\mathrm{tox}} + d \qquad \text{bzw.} \quad Z_{15\%} = d' \cdot \exp\,(c \cdot d_{\mathrm{tox}}) \qquad [4.4]$$

Insgesamt kann man aus der Betrachtung der Abhängigkeit der Zyklenzahl (bei einer prozentualen Abnahme des Programmierfensters) von der Programmierspannung oder der Tunneloxiddicke zwei wichtige Punkte festhalten:

- Bis mindestens 100 °C steigt die Zyklenzahl mit U_{pp} an, bei 250 °C sinkt sie aber
- Der Anstieg (bzw. Abfall) der Zyklenzahl verläuft exponentiell mit U_{pp}

Eine größere Programmierspannung weitet das Programmierfenster (Abbildung 4.5), was bedeutet, dass auch mehr Ladung auf dem Floating-Gate vorhanden ist (Gleichung 4.1). Da bei einem größeren Programmierfenster auch die mögliche Abnahme größer ist, kann also bis zum Erreichen des definierten „Ausfalls" mehr Ladung verloren gehen. Bei 25 °C steigt damit die mögliche Zyklenzahl mit der Programmierspannung an.

Der Zusammenhang zwischen der Programmierspannung und dem Programmierfenster in Abbildung 4.5 ist linear, der Zusammenhang zwischen der Änderung des Programmierfensters und der Ladung auf dem Floating-Gate in Gleichung 4.1 auch, wenn man in erster Näherung von einer konstanten Kapazität zwischen dem Control-Gate und dem Floating-Gate ausgeht. Dennoch ist der Verlauf der Kurven in Abbildung 4.21 nicht linear, sondern exponentiell, also stärker, d. h. es sind mit zunehmender Programmierspannung mehr Zyklen möglich als erwartet. Dies hat mehrere Gründe.

Zum einen zögert die anfängliche Aufweitung des Programmierfensters den Punkt hinaus, an dem das Fenster abnimmt. Dieser Effekt wird mit steigendem U_{pp} noch stärker. Des Weiteren ist der dann folgende Abfall des Fensters nicht linear, sondern im ersten Kurvenabschnitt langsamer und im zweiten Kurvenabschnitt schneller (vergleiche z. B. Abbildung 4.16). Bei 25 °C fällt die Abnahme von 15 % in den langsamer abfallenden Kurvenabschnitt, bei 250 °C eher in den Übergang zum schneller abfallenden Kurvenabschnitt. Je höher die Temperatur ist, desto früher beginnt der Abfall, d. h. mit zunehmender Temperatur liegt eine zusätzliche Defektbildung im Oxid vor. Bei 250 °C dominieren diese Defekte, wodurch das Fenster umso deutlicher sinkt, je größer U_{pp} wird, wodurch immer weniger Zyklen möglich werden und insgesamt die Zyklenzahl mit zunehmender Programmierspannung sinkt.

Für eine Vervollständigung der Analyse wird im folgenden Abschnitt noch der Zusammenhang zwischen der Zyklenzahl und der Temperatur bei einer prozentualen Abnahme des Programmierfensters betrachtet.

Abhängigkeit der „maximalen Zyklenzahl" von der Temperatur

In Abbildung 4.23 ist der Logarithmus der Anzahl der Zyklen nach 15 % Änderung des Programmierfensters in Abhängigkeit von der Temperatur bei zwei verschiedenen Tunneloxiddicken kombiniert mit zwei verschiedenen Programmierspannungen dargestellt. Die zugehörigen Endurance-Kurven befinden sich in Abbildung 4.17.

Die „maximale Zyklenzahl“ nimmt in allen vier Fällen exponentiell mit der Temperatur ab. Die Steigungen der vier Kurven sind sehr ähnlich, die Absolutwerte unterscheiden sich aber. Hierbei fällt wieder auf, dass bis 100 °C die erreichte Zyklenzahl für größere Programmierspannungen auch größer ist, wie es im vorigen Abschnitt diskutiert wurde.

Für die Kombination aus d_{tox} = 13,2 nm und U_{pp} = 16 V gilt der exponentielle Zusammenhang aber nur bis 100 °C. Für größere Temperaturen ändert sich die „maximale Zyklenzahl“ nicht mehr mit der Temperatur, es tritt also eine gewisse Sättigung auf. Dies liegt daran, dass bei d_{tox} = 13,2 nm und U_{pp} = 16 V das elektrische Feld und damit auch das Programmierfenster sehr klein sind, so dass es zu keinem abrupten Zusammenlaufen der Schwellen bei den höheren Temperaturen kommt (vergleiche dazu Abbildung 4.17 c).

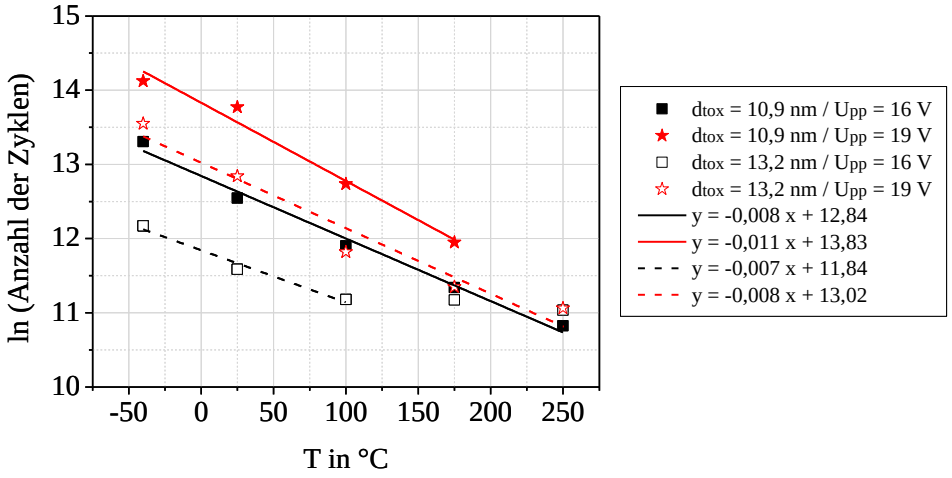


Abbildung 4.23: Logarithmus der Anzahl der Zyklen nach 15 % Änderung des Programmierfensters in Abhängigkeit von der Temperatur; Programmierspannungen U_{pp} = 16 V (Kästchen) und U_{pp} = 19 V (Sterne); Tunneloxiddicken d_{tox} = 10,9 nm (ausgefüllte Symbole) und d_{tox} = 13,2 nm (leere Symbole)

Der Zusammenhang zwischen der Anzahl der Zyklen $Z_{15\%}$ nach 15 % Änderung des Programmierfensters und der Temperatur T in °C kann mit Gleichung 4.5 beschrieben werden:

$$\ln(Z_{15\%}) = -f \cdot T + g \qquad [4.5]$$

Dabei ist f die Steigung der Fit-Geraden (f ist ein positiver Wert) und g der y-Achsenabschnitt. Die Konstanten f und g variieren mit der Programmierspannung und der Tunneloxiddicke. Der Zusammenhang gilt bei höheren Temperaturen (> 100 °C) nicht mehr für kleine Programmierspannungen und dicke Tunneloxide.

Fazit

Aus den Messungen zur Zyklenfestigkeit lassen sich abschließend einige Beobachtungen und Schlussfolgerungen festhalten:

- Der Verlauf des Programmierfensters in Abhängigkeit von der Anzahl der Zyklen hängt von der Temperatur, der Programmierspannung und der Tunneloxiddicke ab.

- Die „maximale Zyklenzahl“ für eine Abnahme von 15 % des anfänglichen Programmierfensters steigt bei 25 °C exponentiell mit der Programmierspannung an, sinkt aber bei 250 °C exponentiell mit der Programmierspannung ab. Entsprechendes gilt für eine kleiner werdende Tunneloxiddicke. Der Übergang von positiver zu negativer Steigung liegt bei etwa 175 °C. Bei einer absoluten Änderung von 1 V sind die Kurven, die die Abhängigkeit der Zyklenzahl von der Programmierspannung bei 25 °C und 250 °C zeigen, parallel.

- Für eine prozentuale Abnahme des anfänglichen Programmierfensters von 15 % sinkt die „maximale Zyklenzahl“ exponentiell mit der Temperatur. Die Steigung und der *y*-Achsenabschnitt hängen von der Programmierspannung und der Tunneloxiddicke ab.

- Die Gründe für das unterschiedliche Temperaturverhalten sind die verschiedenen Größen des Programmierfensters, der unterschiedliche Verlauf mit zunehmender Zyklenzahl und die anfängliche Aufweitung des Fensters, die mit der Programmierspannung zunimmt.

- Bei 250 °C erfolgt eine 15 % - Abnahme des anfänglichen Programmierfensters für die Standardbedingungen (U_{pp} = 16 V und $d_{tox} \approx$ 11 nm) erst nach mehr als 50.000 Zyklen. Eine Änderung von 1 V wird bei diesen Bedingungen schon nach etwas mehr als 30.000 Zyklen erreicht. Nicht mehr voneinander unterscheidbar sind die obere und die untere Schwelle nach etwa 100.000 Zyklen.

- Bei einer Programmierspannung von 16 V und einer Tunneloxiddicke von 11,7 nm sind bei 400 °C immer noch mehr als 1000 Zyklen möglich, bevor die Schwellenspannungen nicht mehr bestimmbar sind.

Kapitel 5

Zuverlässigkeit des Metallisierungssystems

In diesem Kapitel geht es um die Zuverlässigkeit der im untersuchten Hochtemperaturprozess verwendeten Metallisierung. Für den H10-Prozess sind drei Metalllagen Wolfram vorgesehen, die jeweils durch Wolfram-Vias miteinander verbunden werden. Kontakte zum Siliziumfilm und zu der Polysilizium-Ebene spielen bei der Untersuchung ebenso eine Rolle.

In der Literatur findet sich vielfach die Aussage, dass Wolfram sehr elektromigrationsbeständig ist [HAI95], [WER96], [GLA05] und bis zu seiner Rekristallisierungstemperatur von etwa 1800 °C keine durch Elektromigration bedingten Ausfallerscheinungen zeigt [WEN09]. Da Elektromigration ein durch Temperatur beschleunigter Prozess ist, ist die Elektromigrationsbeständigkeit auch der Grund, weshalb statt einer Aluminiummetallisierung bei hohen Betriebstemperaturen Wolfram verwendet wird [CHE95]. Im ersten Abschnitt dieses Kapitels geht es deshalb darum, zu überprüfen, ob bei hohen Stromdichten und einer Betriebstemperatur von 250 °C tatsächlich keine Elektromigration beobachtet werden kann.

Auch wenn Wolfram sehr elektromigrationsbeständig ist, könnten theoretisch auch andere Zuverlässigkeitsprobleme auftreten. Ein wesentlicher Aspekt ist dabei die Stressmigration, die im Gegensatz zur Elektromigration nicht durch eine hohe Stromdichte, sondern durch mechanischen Stress ausgelöst wird. Zuverlässigkeitsuntersuchungen zur Stressmigration werden im zweiten Abschnitt diskutiert.

5.1 Elektromigration

5.1.1 Theoretische Aspekte der Elektromigration

Unter dem Begriff Elektromigration versteht man die durch eine hohe Stromdichte hervorgerufene Wanderung von Material. Sie ist zu unterscheiden von der Thermomigration, der Materialwanderung aufgrund eines Temperaturgradienten und von der Stressmigration, der Materialwanderung infolge mechanischen Stresses. Beide können aber den Materialfluss der Elektromigration noch verstärken [BER06].

Jede Metallebene in einem mikroelektronischen Prozess besteht aus einer polykristallinen Metallschicht, d.h. sie ist aus Kristallkörnern gleicher Kristallstruktur aber unterschiedlicher Orientierung aufgebaut. Der Übergang zwischen den Kristallkörnern wird Korngrenze genannt. Fließt ein Strom durch das Metall, so kommt es zu einer Wechselwirkung der bewegten Elektronen mit dem Kristallgitter. Die Elektronen übertragen einen Teil ihres Impulses an die Metallionen des Gitters. Vor allem die an den Korngrenzen schwach gebundenen Metallionen werden aus ihrer Struktur herausgelöst und wandern bevorzugt entlang der Korngrenzen in Richtung des Elektronenflusses [LIE02].

Zwei Auswirkungen dieser Materialwanderung können beobachtet werden. Dort, wo ein Materialabtrag stattfindet, bilden sich Hohlräume (engl. voids), die zu einer offenen Stelle in der Metallisierung führen können bzw. den Widerstand der Bahn, des Vias oder des Kontakts stark anwachsen lassen. An der Stelle, zu der die Metallionen getragen werden, sammelt sich Material an. Diese Ansammlungen (engl. hillocks) können so groß werden, dass ein Kurzschluss zu benachbarten Leiterbahnen möglich wird [WIT00].

Wie anfällig eine Metallisierung für Elektromigration ist, hängt von vielen Faktoren ab, vor allem von den Eigenschaften des verwendeten Metalls und dessen Prozessierung. Die Korngröße und die Verteilung der Körner haben dabei einen entscheidenden Einfluss. In Bereichen, in denen ein Übergang von kleineren Körnern zu größeren Körnern zu finden ist, kann das abgetragene Material besser zu- als abfließen und es kommt verstärkt zur Bildung von Hohlräumen. In Gebieten, in denen größere zu kleineren Körnern übergehen, treten vor allem Materialansammlungen auf [LIE02].

Eine weitere Rolle spielt auch der Verlauf der Leiterbahnen bzw. die Anordnung von Vias und Kontakten, insbesondere bei Abknickungen von Bahnen. Dabei sind 90°-Winkel besonders anfällig für Elektromigration, da hier der Strom nur durch einen Teil der Ecke fließt, die dann stärker beschädigt werden kann [LIE06].

Auch die Geometrie der Metallverbindungen ist entscheidend. Je schmaler, länger und dünner eine Bahn ist, desto größer ist ihre Anfälligkeit für Elektromigration. Dabei gibt es aber eine Grenze, unterhalb der die Elektromigrationsbeständigkeit wieder zunimmt. Wenn die Bahnbreite kleiner ist als die mittlere Korngröße des Metalls, liegen zwangsläufig viele Korngrenzen quer zum Stromfluss und nur wenige Korngrenzen längs der Flussrichtung des Stromes. Der Materialtransport in Richtung des Stromflusses wird somit erschwert bzw. verhindert und die Leiterbahn ist besser vor Elektromigration geschützt. Eine solche Struktur wird Bambusstruktur genannt [WAL92].

Neben der Breite der Leiterbahn ist auch die Länge entscheidend. Blech *et al*. fanden heraus, dass es für eine bestimmte Stromdichte eine Bahnlänge gibt, unterhalb der keine Elektromigration auftritt [BLE76]. Der Grund dafür ist ein gewisser Materialrückfluss, wenn der Stressgradient, der sich bei einer langen Metallbahn einstellt, einen kritischen Wert unterschreitet. Diese Länge wird Blech-Länge genannt.

Ein weiterer Aspekt ist zudem die Abhängigkeit der Elektromigration von der Stromrichtung. Bei Wechselstrombetrieb ist die Metallisierung weniger anfällig als bei Gleichstrombetrieb, da der in Stromrichtung wirkende Materialtransport ebenso seine Richtung ändert und so Materialansammlungen bzw. Materialabtragungen kompensiert werden [OSH12], [LIE02].

Elektromigration betrifft vor allem Aluminium und Aluminiumverbindungen. Kupfer ist mit einem Schmelzpunkt von 1084,62 °C [HCP11] weniger anfällig für Elektromigration und zudem ein besserer Leiter als Aluminium (Schmelzpunkt bei 660,323 °C unter Normaldruck [HCP11]), aber prozesstechnisch schwieriger zu handhaben. Oft werden deshalb Aluminium-Kupfer-Verbindungen (Aluminium mit circa 0,5 % Kupfer) als Leiterbahnmaterial eingesetzt [AME70], [TU03]. Bei modernen Prozessen werden auch Kupfermetallisierungen verwendet. Bei reinem Kupfer erfolgt die Materialwanderung hauptsächlich an den Grenzflächen zwischen den Metallebenen und dem umgebenden Oxid [GLA05].

Für Hochtemperaturprozesse ist Wolfram wegen seiner Elektromigrationsbeständigkeit die beste Wahl. Wolfram ist ein Metall mit einem sehr hohen Schmelzpunkt von 3414 °C bei

Normaldruck und einer großen Dichte von 17,6 g/cm^3 bei 20 °C. Bei Temperaturen, wie sie in der Siliziumtechnologie möglich sind, ist deshalb keine Materialdiffusion zu erwarten bzw. diese so klein, dass sie nicht messbar ist. Erst ab Temperaturen oberhalb von 1600 °C [OBO65] bzw. 1800 °C [WEN09] kann in Wolframleitern Materialwanderung beobachtet werden. Dies spielt für die Mikroelektronik keine Rolle, ist aber zum Beispiel für die Glühwendel in Glühlampen von Bedeutung [OBO65], [PEA68], [OSH12].

Andererseits hat Wolfram bei 20 °C mit ungefähr $5{,}28 \cdot 10^{-8}$ Ω·m gegenüber Aluminium (circa $2{,}65 \cdot 10^{-8}$ Ω·m) einen zwei- bis dreimal größeren spezifischen Widerstand [HCP11]. Das genaue Verhältnis hängt von der jeweiligen kristallinen Beschaffenheit der Metallschicht ab. Der höhere Widerstand ist der wesentliche Nachteil der Wolframmetallisierung und der Grund, weshalb bei Nicht-Hochtemperaturprozessen Wolfram als Leiterbahnmaterial eher selten verwendet wird.

5.1.2 Messmethode der Elektromigration

Im JEDEC-Dokument JP001 [JED1] finden sich Erläuterungen, wie Elektromigration von Metallbahnen, Vias und Kontakten zu messen ist und wie eine Lebensdauer für die Metallisierung berechnet werden kann. Weitere relevante JEDEC-Standards sind JESD61, JESD63 und JESD87 [JED61], [JED63], [JED87].

Das Modell, auf dem die Lebensdauervorhersagen infolge von Elektromigration beruhen, wurde von Black entwickelt [BLA67], [BLA69]. Die Blacksche Gleichung beschreibt die Lebensdauer t_{fail} als Zusammenhang zwischen der Stromdichte j und der Temperatur T:

$$t_{\mathrm{fail}} = A \cdot \left(\frac{1}{j^n} \right) \cdot \exp\left(\frac{E_{\mathrm{A}}}{k \cdot T} \right) \qquad [5.1]$$

Hierbei ist A eine material- und geometrieabhängige Konstante, n eine ebenso materialabhängige Konstante, E_{A} die Aktivierungsenergie für Elektromigration und k die Boltzmann-Konstante. Die Blacksche Gleichung ist ein empirisches Modell. Der exponentielle Zusammenhang zeigt, dass die Temperatur ein kritischer Faktor bei der Elektromigration ist. Für eine genaue Lebensdauervorhersage ist die Kenntnis der exakten Temperatur deshalb sehr wichtig. Die zweite Größe, die Elektromigration beeinflusst, ist die Stromdichte. Je höher die Temperatur und die Stromdichte sind, desto geringer ist die Lebensdauer.

Ziel der in den JEDEC-Dokumenten beschriebenen Messmethode ist es, für die festgelegte maximale Stromdichte der Technologie und die maximale Betriebstemperatur die Lebensdauer der Metallisierung anhand der Blackschen Gleichung berechnen zu können. Dazu wird die Teststruktur auf eine konstante Temperatur geheizt und mit einem konstanten Strom belastet, der einer Stromdichte oberhalb der maximal erlaubten entspricht. Die Widerstandsänderung dieser Struktur wird über der Zeit beobachtet. Die Zeit, bei der der Widerstand eine festgelegte prozentuale Änderung, z.B. von 2 %, 10 % oder 100 %, erfahren hat, oder bei der ein Kurzschluss auftritt, ist die Ausfallzeit t_{fail} für die verwendete Temperatur-Strom-Kombination. Aus Messungen der Ausfallzeiten bei der gleichen Stromdichte aber verschiedenen Temperaturen kann die Aktivierungsenergie E_{A}, aus Messungen bei derselben Temperatur und unterschiedlichen Stromdichten der Exponent n bestimmt werden. Die durch Elektromigration bedingten Ausfallzeiten sollten einer log-normalen Verteilung folgen [GAL01].

Wenn innerhalb einer gewissen Zeit eine gleichmäßige Änderung des Widerstandes erfolgt, kann statt die Ausfallzeit zu messen auch auf diese extrapoliert werden. Für Aluminiummetallisierungen, bei denen aus prozesstechnischen Gründen zusätzliche Metallebenen wie beispielsweise aus Titannitrid (TiN) ober- oder unterhalb der Aluminiumlage aufgebracht wurden, kann beispielsweise ein linear mit der Zeit zunehmender Widerstand beobachtet werden. In diesem Fall wäre eine Extrapolation auf die Ausfallzeit möglich. Wenn ein Widerstand aber nicht gleichmäßig mit der Zeit zunimmt, sondern nahezu konstant bleibt und zu einem bestimmten Zeitpunkt eine abrupte Widerstandsänderung erfährt, ist eine Extrapolation nicht möglich. Dies ist zum Beispiel bei reinen Aluminiummetallisierungen der Fall.

Die Aktivierungsenergie für die strombedingte Materialdiffusion im Leiter ist abhängig vom verwendeten Metall. Für eine reine Aluminiummetallisierung ergeben sich Werte zwischen 0,6 eV [BER06] und 0,95 eV [JED122], wobei es darauf ankommt, ob Diffusion hauptsächlich an den Korngrenzen (0,68 eV nach [JED122]), wie es meistens der Fall ist, oder an den Grenzflächen zu den umgebenden Materialien (0,95 eV nach [JED122]) stattfindet. Handelt es sich um eine Aluminium-Kupfer-Verbindung (Aluminium mit circa 0,5 % Kupfer) oder eine reine Kupfermetallisierung liegt die Aktivierungsenergie bei etwa 0,9 eV [JED122]. Auch für die Konstante n ergeben sich verschiedene Werte. Black nahm für Aluminium $n \approx 2$ an, aber der Wert hängt vom Material und der Stromdichte ab [BER06]. Für Wolfram findet man in der Literatur wegen der hohen Elektromigrationsbeständigkeit keine Angaben.

5.1.3 Elektromigrationstests an Teststrukturen

Die Elektromigrationstests für die H10-Technologie wurden an Kelvin-Strukturen aus Metallbahnen, Metallmäandern, Vias und Kontakten durchgeführt. Die Strukturen wurden nach den Grundlagen der JEDEC-Dokumente angefertigt (siehe Anhang B).

In Abbildung 5.1 ist eine Rasterelektronenmikroskopaufnahme des H10-Backends dargestellt. Es besteht aus drei etwa 0,55 µm dicken Wolframlagen, die durch Chemische Gasphasenabscheidung (Chemical Vapor Deposition, kurz CVD) abgeschieden werden und durch Oxidschichten voneinander getrennt sind. Unter jeder Metallebene befindet sich eine Schicht aus Titan und Titannitrid, da eine direkte Abscheidung von Wolfram auf Oxid prozesstechnisch schwierig ist. Weiterhin sind drei Aktivgebietskontakte aus Wolfram zwischen der ersten Metallebene und dem Siliziumfilm zu erkennen. Zur Passivierung sind eine Oxid- und dann eine Nitridschicht aufgebracht. Kontakte zwischen der ersten Metalllage und Polysilizium sowie Vias zwischen der ersten und zweiten Metallebene (Via 1) bzw. der zweiten und dritten Metallebene (Via 2) sind in diesem Ausschnitt nicht dargestellt.

Während der Entwicklung des H10-Prozesses wurden Werte für die maximalen Stromdichten von Metallbahnen, Vias und Kontakten festgelegt. Für eine Bahn aus Metall 1, Metall 2 oder Metall 3 ist eine maximale Stromdichte von 0,6 mA pro µm Bahnbreite erlaubt. Bei einer Schichtdicke von 0,55 µm bedeutet dies eine querschnittsflächenbezogene Stromdichte von 0,11 MA/cm². Für Kontakte und Vias wurden maximal 0,5 mA pro Kontaktloch bzw. pro Via festgelegt, d. h. für 0,8 µm x 0,8 µm große Via- bzw. Kontaktlöcher sind maximal 0,08 MA/cm^2 erlaubt.

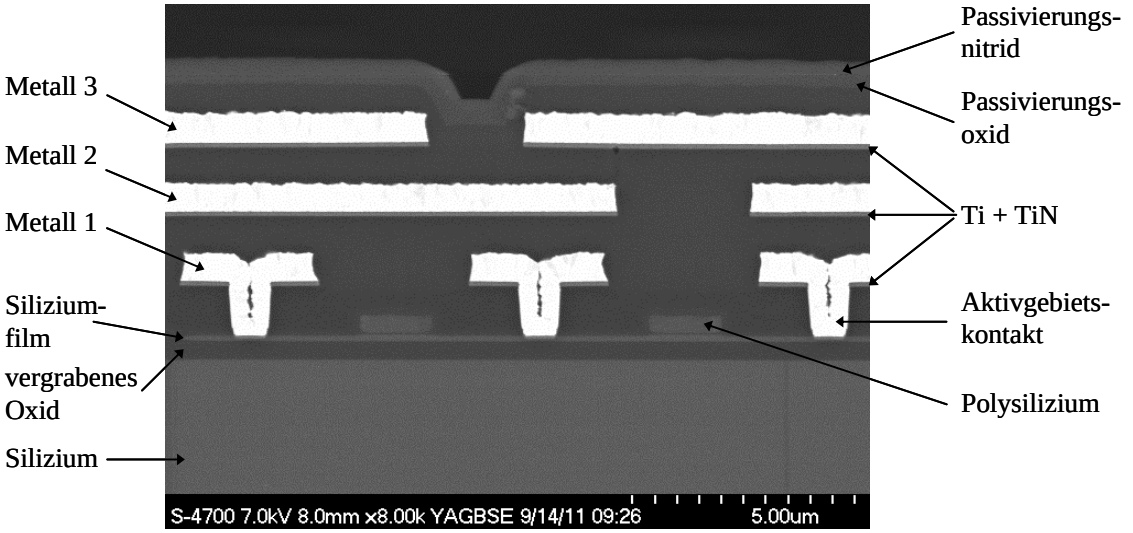


Abbildung 5.1: Querschnitt des H10-Backends als Rasterelektronenmikroskopaufnahme mit drei Metallebenen aus Wolfram und drei Aktivgebietskontakten

Um eine Lebensdauer berechnen zu können, ist es notwendig, für eine gegebene Stressbedingung aus Stressstromstärke und Stresstemperatur die Ausfallzeit zu kennen. Dies kann, wie in Abschnitt 5.1.2 beschrieben, entweder durch Messen dieser Zeit oder durch eine Extrapolation geschehen. Dafür muss aber der Verlauf der Widerstandsänderung mit der Zeit eindeutig sein. Für die durchgeführten Elektromigrationsmessungen an verschieden langen und breiten Metallbahnen aller drei Metalllagen bei Temperaturen bis 300 °C und Stromdichten bis etwa 2,7 MA/cm^2 war weder eine Widerstandsänderung in einer akzeptablen Zeit messbar noch ein klarer Verlauf sichtbar[1]. Abbildung 5.2 zeigt als Beispiel die zeitabhängige Widerstandsentwicklung an einer 7000 µm langen Metallbahn aus Metall 2.

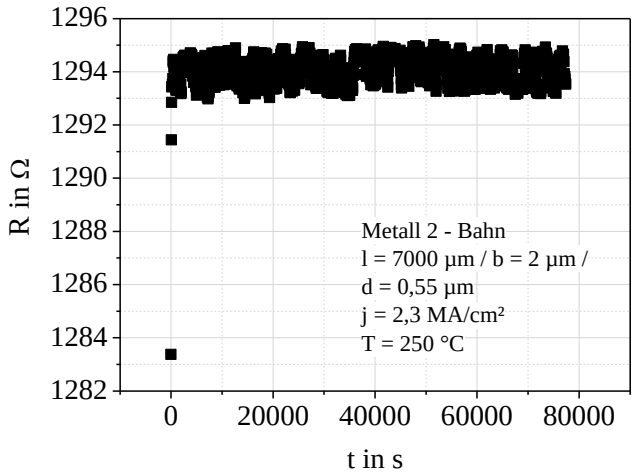


Abbildung 5.2: Widerstand einer 7000 µm langen Metall 2 - Bahn aus Wolfram in Abhängigkeit von der Zeit bei einem konstant angelegten Strom von 25 mA; mit einer Bahnbreite von b = 2 µm und einer Schichtdicke von d = 0,55 µm ergibt sich eine Stromdichte von j = 2,3 MA/cm^2; Temperatur des Thermochucks: T = 250 °C; Kontaktierung über Pads mit einer zusätzlichen Aluminiumschicht

[1] Die maximal mögliche Stromdichte ist begrenzt, da für eine Messnadel nur eine maximale Stromstärke von 30 mA zugelassen ist.

In den ersten Sekunden erwärmt sich die Teststruktur auf etwa 255 °C, wodurch der Widerstand ansteigt, danach bleibt er aber auf einem konstanten Niveau. Die Schwankung um ± 1 Ω ist auf die Temperaturschwankung des Thermochucks zurückzuführen. Gemessen wurde bei der maximalen Betriebstemperatur von 250 °C mit einer circa 21-mal so großen Stromdichte wie im Betrieb maximal erlaubt.

Auch Elektromigrationstests an Vias zeigten keine Widerstandsdegradationen. Die Kontaktierung erfolgte dabei immer über Pads mit einer zusätzlichen Aluminiumschicht. Dadurch war der Kontakt zwischen den Nadeln und den Pads auch bei einer Wafertemperatur von 250 °C bzw. 255 °C stabil.

Abbildung 5.3 zeigt den zeitlichen Verlauf der Widerstände einzelner Kontakte zwischen dem n-dotierten Polysilizium und Metall 1, die bei 250 °C mit einem konstanten Strom von 15 mA belastet wurden. Die Stromdichte ist damit um einen Faktor 30 höher als im Betrieb maximal erlaubt und genauso groß wie im Falle des Metallbahntests in Abbildung 5.2.

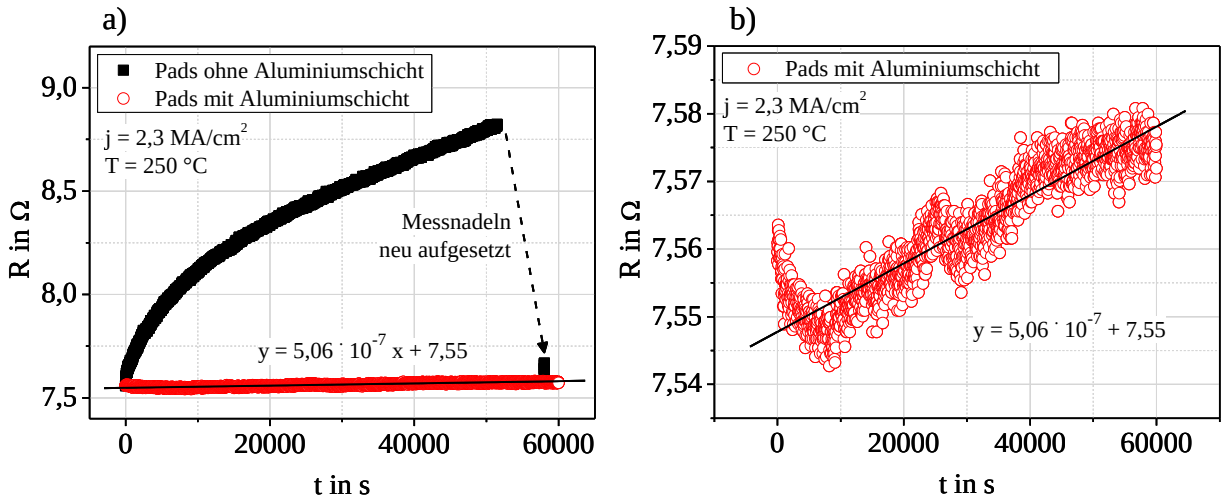


Abbildung 5.3: Widerstände von Polysilizium-Metall-Einzelkontakten aus Wolfram in Abhängigkeit von der Zeit bei einem konstanten Strom von 15 mA; mit einer Kontaktlänge und Kontaktbreite von $l = b = 0{,}8$ µm ergibt sich eine Stromdichte von $j = 2{,}3$ MA/cm²; Temperatur des Thermochucks: $T = 250$ °C; Kontaktierung über Pads ohne einer zusätzlichen Aluminiumschicht (schwarze Kästchen in a)) bzw. über Pads mit einer zusätzlichen Aluminiumschicht (rote leere Kreise in a) und b)); die Messwerte, die der Kurve der schwarzen Kästchen zugrunde liegen (inklusive dem Messwert nach dem erneuten Aufsetzen der Nadeln), sind aus Darstellungsgründen alle um einen Offset von 1,8 Ω nach unten verschoben worden

Bei den Untersuchungen an Kontaktwiderständen konnten zwei Aspekte festgestellt werden. Wenn auf den vier Pads zur Kontaktierung der Messnadeln keine zusätzliche Aluminiumschicht vorhanden ist, steigt der gemessene Widerstand mit der Zeit an (schwarze Kästchen in Abbildung 5.3 a)). Bei einem erneuten Aufsetzen der Messnadeln kann die Widerstandsänderung aber wieder rückgängig gemacht werden. Der Grund für den Anstieg ist die Degradation des Messnadel-Pad-Kontaktes. Bei 250 °C oxidiert die Wolframoberfläche der Pads und bildet einen zusätzlichen Widerstand. Bei einem erneuten Aufsetzen der Nadeln wird die gebildete Oxidschicht wieder weggekratzt. Für Messstrukturen, die an Pads mit einer zusätzlichen Aluminiummetallisierung angeschlossen sind, wie es auch bei den Metallbahn- und Viastrukturen der Fall war, kann diese große Widerstandsänderung nicht festgestellt werden (rote leere Kreise in Abbildung 5.3 a)).

Im Gegensatz zu den Metallbahn- und Viastrukturen ist aber möglicherweise eine geringe, wenn auch nicht signifikante Änderung mit der Zeit messbar (rote leere Kreise in Abbildung 5.3 b)[2]. Die Messpunkte steigen zwar insgesamt an, schwanken aber auch als Folge der Temperaturregelung des Thermochucks. Extrapoliert man dennoch unter der Annahme von $n = 2$ (siehe Gleichung 5.1) die Änderung des Widerstands aus Abbildung 5.3 b) auf Betriebsbedingungen (T = 250 °C und j = 0,5 mA pro Kontaktloch), so wäre eine Widerstandsänderung von 10 % erst nach mehr als 42 Jahren zu erwarten.

Ob es sich also tatsächlich um eine durch Elektromigration bedingte Degradation handelt, ist nicht sicher. Zudem ist auch eine leichte Oxidation der mit Aluminium beschichteten Pads nicht auszuschließen. Um jeglichen Einfluss der Oxidation, vor allem der Pads, zu vermeiden, müssten entweder alle Elektromigrationstests unter Schutzgas durchgeführt werden oder die Pads mit einer nicht-oxidierbaren Schicht wie zum Beispiel Gold versehen sein. Da die hier beobachtete Degradation aber sehr gering ist und außerdem andere Probleme bei der Zuverlässigkeit der Polysilizium-Metall-Kontakte auftraten (siehe Abschnitt 5.2.3), wurde keine umfangreiche Statistik zur Elektromigration von Kontakten erstellt.

Alle Elektromigrationstests wurden an passivierten Wafern durchgeführt. In früheren Untersuchungen zeigte sich, dass Widerstandsänderungen der Wolframmetallisierung bei nicht-passivierten Wafern auf die Oxidation der Leiterbahnen zurückzuführen sind [WER96].

Fazit

Zusammenfassend kann man festhalten, dass Elektromigration bei einer Wolframmetallisierung bei 250 °C tatsächlich keine Rolle spielt, wie es auch in der Literatur beschrieben wird. Bei Kontakten sind nur geringe, nicht signifikante Änderungen festzustellen, die aber auch andere Ursachen als Elektromigration haben können. Starke Widerstandszunahmen wie in Abbildung 5.3 a) sind immer auf die Degradation des Messnadel-Pad-Kontaktes durch die Oxidation von Pads, die keine zusätzliche Aluminiumschicht haben, zurückzuführen. Oxidationen der Nadeln spielen keine Rolle.

[2] Die Kurven in Abbildung 5.3 a) und b), die aus leeren Kreisen bestehen, sind identisch.

5.2 Stressmigration

5.2.1 Theoretische Aspekte der Stressmigration

Stressmigration bezeichnet das Entstehen von Hohlräumen (engl. voids) in Leitern als Folge einer Materialwanderung, die durch thermo-mechanischen Stress zwischen der Metallisierung und den sie umgebenden Oxidbereichen hervorgerufen bzw. beeinflusst wird. Dieser Effekt wird auch als Stress Voiding oder als Stress-Induced Voiding [JED139] bezeichnet und wurde erstmals 1984 von Curry *et al.* beschrieben [CUR84]. Im Gegensatz zur Elektromigration ist es nicht der Stromfluss, der für die Materialwanderung durch den Leiter verantwortlich ist, sondern mechanischer Stress als Folge einer Temperaturänderung, der Risse im Metall erzeugt, die dem anliegenden Stress entgegenwirken. Der Grund für den Stress ist der unterschiedliche thermische Ausdehnungskoeffizient der Materialien. Wolfram besitzt beispielsweise einen thermischen linearen Ausdehnungskoeffizienten von $4{,}5 \cdot 10^{-6}$ pro K [HCP11], Siliziumdioxid hingegen hat einen Koeffizienten von ungefähr $0{,}5 \cdot 10^{-6}$ pro K [VIR02].

Stressmigration hängt aber nicht allein davon ab, wie groß der thermo-mechanische Stress ist, sondern auch von der Möglichkeit, dass Material überhaupt im Leiter wandern kann. Dabei ist zu beachten, dass die beiden Phänomene, die Stressmigration ausmachen, nämlich die Materialdiffusion und der thermo-mechanische Stress, unterschiedlich temperaturabhängig sind.

Hohlräume entstehen durch eine Abwanderung von Material bzw. durch eine Zuwanderung von Leerstellen. Die Diffusivität D dieser Leerstellen im Metall nimmt mit der Temperatur T exponentiell zu [GLA05], [SUL92]:

$$D = D_0 \cdot \exp\left(\frac{-E_A}{k \cdot T}\right) \qquad [5.2]$$

D_0 ist eine Konstante, E_A die Aktivierungsenergie für die Leerstellendiffusivität und k die Boltzmann-Konstante. Gleichung 5.2 zeigt, dass mit steigender Temperatur die Hohlraumbildung erleichtert wird. Dabei ist sowohl eine Diffusion an den Grenzflächen zwischen Metall und umgebendem Material [GLE97] als auch eine Diffusion im Leiter entlang der Korngrenzen möglich [OVI08].

Der thermisch bedingte mechanische Stress, auch thermo-mechanischer Stress genannt, nimmt bei der Abkühlung des Wafers nach einem Back-End-Prozessschritt, der bei einer erhöhten Temperatur ausgeführt wurde, zu. Je stärker der Wafer abgekühlt wird, d.h. je weiter sich die Temperatur von der Abscheidetemperatur entfernt, desto größer wird der mechanische Stress, desto kleiner wird aber die Diffusivität der Leerstellen.

Die beiden Phänomene hängen zwar voneinander ab, wirken einander aber auch entgegen, so dass es eine Temperatur gibt, bei der der Effekt der Stressmigration am kleinsten ist. Deshalb findet auch das Arrheniusgesetz keine Anwendung (vergleiche dazu auch Gleichung 5.3 im nächsten Abschnitt).

Die sich bildenden Hohlräume führen zu einer Zunahme des Widerstands und können die Leiterbahn sogar komplett öffnen. Da in hohlen Bereichen eines Leiters die lokale Stromdichte zunimmt, kann Stressmigration das Phänomen der Elektromigration verstärken (siehe Abschnitt 5.1.1).

Stressmigration wirkt sich stärker aus, je dünner oder schmaler die Metallbahn ist, da hier von der Grenzfläche wachsende Hohlräume schneller zu einer offenen Leiterbahn führen. Aus diesem Grund sind Vias und Kontakte besonders anfällig für Stressmigration.

Um Stress schon während der Prozessierung zu vermeiden, sollten die Abscheidetemperaturen im Backend so gering wie möglich sein. Da Wolfram eine höhere Abscheidetemperatur hat als Aluminium, ist es auch anfälliger für thermisch bedingten mechanischen Stress bei der Abkühlung. Obwohl der Stress bei der Prozessierung eines Wolfram-Back-Ends höher ist als bei der Prozessierung eines Aluminium-Back-Ends [AHN87], [ITO91], ist die Anfälligkeit von Wolfram für die Bildung von Hohlräumen geringer als beispielsweise bei Aluminium, da infolge des hohen Schmelzpunktes von Wolfram die Diffusivität kleiner ist. Der prozessbedingte Stress bei Wolfram muss also nicht zwangsläufig zur Degradation führen, kann aber die Mikrostruktur der Metallbahnen ändern und damit auf die Stressmigration einwirken [SUN11].

Auch wenn sich die meisten Untersuchungen in der Literatur auf Aluminium [TEZ90], [OKA93], [GLE97] oder Kupfer [MIC01], [FIS02], [GLA05] beziehen, ist für den H10-Prozess eine Widerstandsänderung der Metallisierung durch Stressmigration nicht auszuschließen, weshalb entsprechende Untersuchungen sinnvoll sind.

5.2.2 Messmethode der Stressmigration

Die im Folgenden beschriebene und für den H10-Prozess angewandte Methode ist im JEDEC-Standard JEP139 [JED139] und im JEDEC-Dokument JP001 [JED1] beschrieben. Sie bezieht sich auf eine Aluminiummetallisierung, weshalb im Falle von Wolfram Abweichungen möglich sind. Über die Erstellung passender Teststrukturen gibt der Standard JESD87 [JED87] Auskunft.

Zur Berechnung der Lebensdauer von Metallbahnen oder Vias werden zunächst die Widerstände bei Raumtemperatur vermessen[3]. Danach werden die Strukturen bei verschiedenen Temperaturen ohne Strombelastung gelagert. Nach gewissen Zeitabständen werden die Wafer aus dem Ofen geholt und die Widerstände der Strukturen wieder bei Raumtemperatur gemessen. Das Ausfallkriterium ist eine zuvor festgelegte prozentuale Widerstandsänderung. Laut JEDEC-Dokument JP001 sind die Ausfallzeiten log-normal verteilt.

Nach dem JEDEC-Standard JED139 ergibt sich damit in Anlehnung an die Blacksche Gleichung für die Ausfallzeit bei einer Lagerungstemperatur T:

$$t_{\text{fail}} = A \cdot \left(\frac{1}{\Delta T^2} \right) \cdot \exp\left(\frac{E_{\text{A}}}{k \cdot T} \right) \qquad [5.3]$$

A ist eine material- und geometrieabhängige Konstante, ΔT die Differenz zwischen der Abscheidetemperatur der Passivierung und der Lagerungstemperatur, E_{A} die Aktivierungsenergie für Stressmigration und k die Boltzmann-Konstante. Für den H10-Prozess ist die

[3] Im Gegensatz zu den Messvorschriften zur Elektromigration werden im JEDEC-Standard JEP139 nur Stressmigrationsmessungen an Metallbahnen und Viaketten vorgeschrieben. Kontakte werden nicht explizit erwähnt.

Abscheidetemperatur der Nitrid-Passivierung mit 417 °C die höchste Temperatur während der Back-End-Prozessierung[4].

Der Exponent der Temperaturdifferenz ΔT ist im JEDEC-Standard JED139 mit 2 angegeben und bezieht sich auf weiche Metalle wie Aluminium oder Kupfer. Für härtere Metalle wie zum Beispiel Wolfram kann er bis zu 9 betragen [McP10]. Für die Aktivierungsenergien finden sich im JEDEC-Dokument JEP122E die Angaben von 0,6 eV im Fall von Aluminium und 0,74 eV bis 1,2 eV für Kupfermetallisierungen [JED122]. Durch den Fokus auf Aluminium und Kupfer sind wie auch bei der Elektromigration Angaben für Wolfram in der Literatur nicht zu finden.

5.2.3 Stressmigrationstests an Teststrukturen

Wie vom JEDEC-Standard JEP139 vorgeschrieben, wurden Stressmigrationstests an Metallbahnen, Metallmäandern und Viaketten auf passivierten Wafern durchgeführt. Abbildungen der Teststrukturen befinden sich in Anhang B. Die Widerstände der einzelnen Teststrukturen wurden nach dem Kelvin-Prinzip bei 25 °C gemessen und die Wafer danach bei 250 °C und 350 °C gelagert. Obwohl in den JEDEC-Standards nicht vorgeschrieben, wurde zusätzlich auch das Verhalten von Kontaktwiderständen nach Temperaturlagerung untersucht.

Metallbahn-Strukturen

Abbildung 5.4 zeigt die Auswirkung der Temperatur-Lagerung auf die Widerstände[5] von 7000 µm langen Metallbahnen aus Metall 1, Metall 2 und Metall 3. Die Diagramme (a), (c) und (e) auf der linken Seite beinhalten die Ergebnisse der 250 °C-Lagerung, die Diagramme (b), (d) und (f) auf der rechten Seite die Resultate der 350 °C-Lagerung. Für jede Metalllage sind bei beiden Temperaturen jeweils drei Bahnen gelagert worden (Kästchen, Dreiecke, Sterne). Die Messwerte vor der Lagerung (t = 0 h) sind wegen der logarithmischen x-Achsendarstellung bei t = 1 h eingezeichnet[6].

Nach mehr als 10.000 h bei 250 °C haben sich die Widerstände der Metallbahnen um weniger als 1 % erhöht. Die kleinen Schwankungen zwischen den einzelnen Messwerten sind darauf zurückzuführen, dass die Raumtemperatur, bei der die Widerstände ausgelesen wurden, von Auslesezeitpunkt zu Auslesezeitpunkt leicht variierte. Dasselbe Ergebnis wurde an verschiedenen Metallbahnmäandern erzielt.

Im Vergleich zur 250 °C-Lagerung waren die Widerstände der Metallbahnen bei der 350 °C-Lagerung nur bis etwas mehr als 2000 h stabil. Einige Teststrukturen fielen sogar schon früher aus[7]. Danach lagen in allen Fällen Widerstandszunahmen von mehr als 100 % vor. Auch für verschiedene Metallmäander erhielt man vergleichbare Ergebnisse.

[4] Anschließend folgt noch die Abschlusstemperung bei 450 °C, die für die Stressmigration aber nicht relevant ist.
[5] Die Widerstände wurden jeweils aus den erhaltenen Spannungswerten bei Raumtemperatur für den angelegten Strombereich von 0,5 mA bis 7,5 mA berechnet.
[6] Auch bei den folgenden Abbildungen werden die Messwerte von t = 0 h bei t = 1 h eingezeichnet, sofern eine logarithmische x-Achse vorliegt.
[7] Für eine bessere Übersichtlichkeit sind in Abbildung 5.4 nur die Messwerte vor dem Ausfall der Teststrukturen eingezeichnet.

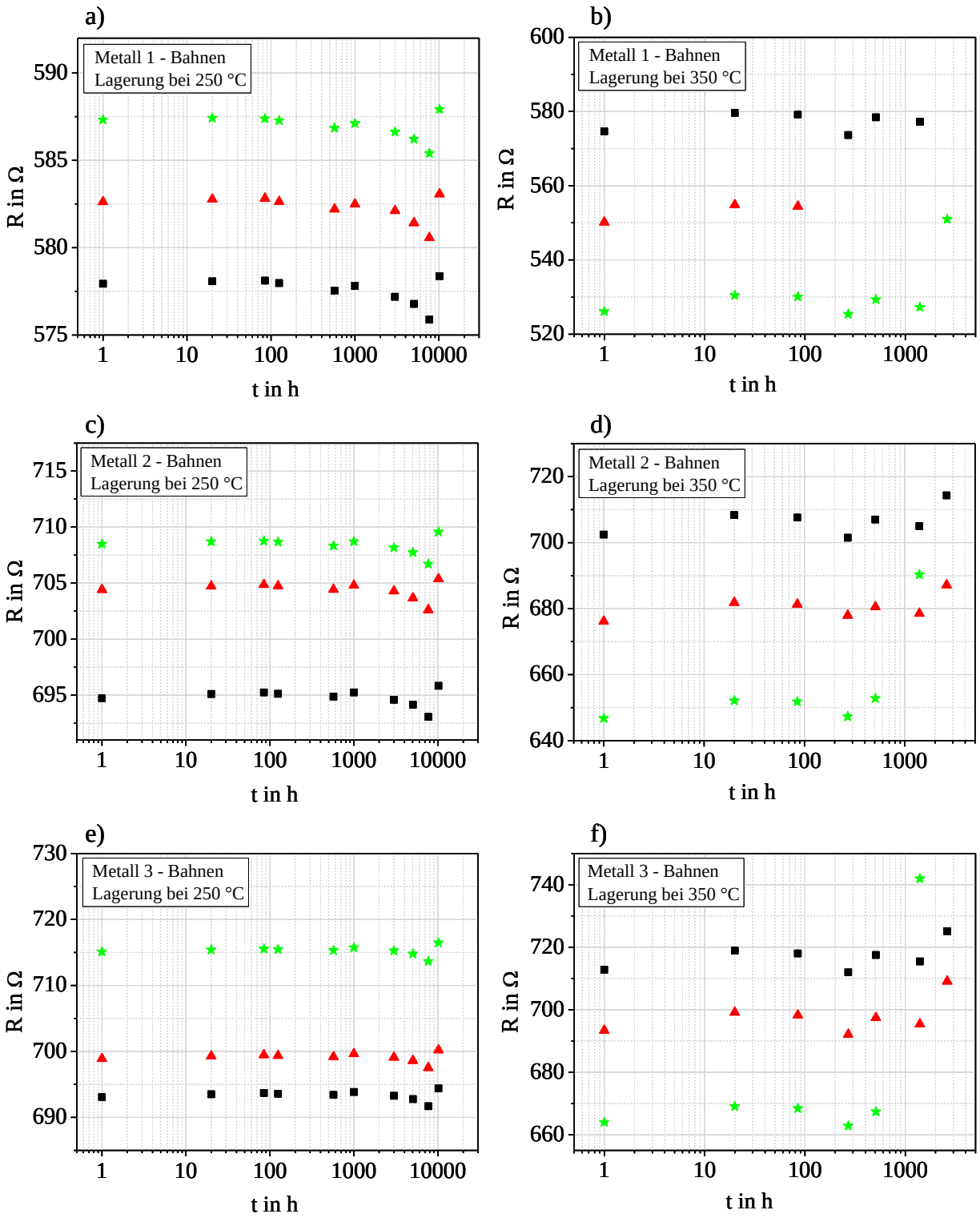


Abbildung 5.4: Widerstände von jeweils drei Metallbahnen aus Wolfram in Abhängigkeit von der Zeit bei einer Lagerungstemperatur von 250 °C (a, c, e) bzw. 350 °C (b, d, f); die Werte vor der Lagerung bei t = 0 h sind wegen der logarithmischen x-Achsendarstellung bei t = 1 h eingezeichnet; Bahnlänge l = 7000 µm, Bahnbreite b = 2 µm, Schichtdicke d = 0,55 µm

Die Ausfälle der Metallbahnen bei 350 °C sind auf Oxidationen im Bereich der Padmetallisierung zurückzuführen und keine Folge von Stressmigration. Die Pads, die die Metallbahnen anschließen, bestehen aus einem Stapel aller drei Wolfram-Ebenen und beider Vias. Hinzu kommt eine abschließende Aluminiummetallisierung zur besseren Kontaktierung der Mess-

nadeln. Zu diesem Zweck ist auf ihnen die Passivierung geöffnet. Kratzt man beim Messen mit den Nadeln an der Aluminiumoberfläche, legt man auch den Zugang zur obersten Wolframebene frei, die dann im Ofen oxidieren kann[8]. Nach 1400 h bei 350 °C zeigten die meisten Pads schon einige farbliche Veränderungen und schwarze Bereiche bei der Überleitung von den Pads zu den Metallbahnen (Abbildung 5.5 a)). Allerdings waren auch einige, wenn auch wenige, unkontaktierte Pads betroffen (Abbildung 5.5 b)). Je länger die Wafer im Ofen lagen, desto größer wurden die Ausmaße der Oxidationen. An Pads von Randstrukturen konnten sogar schon nach weniger als 100 h bei 350 °C Oxidationen beobachtet werden (z. B. bei der einen Metall 1 - Bahn aus Abbildung 5.4 b)). Bei 450 °C war schon nach 1,5 h eine deutliche Veränderung der Pads zu erkennen (Abbildung 5.5 c)), weshalb Langzeitlagerungen bei 450 °C nicht möglich waren. Nach mehr als 10.000 h bei 250 °C gab es keinerlei Veränderungen an den Pads (Abbildung 5.5 d)).

Vermutlich sind also die Metallbahnen auch nach 2000 h bei 350 °C noch intakt, ihr Widerstand kann aber nicht gemessen werden, da kein Kontakt mehr zwischen den Nadeln und den Metallebenen möglich ist.

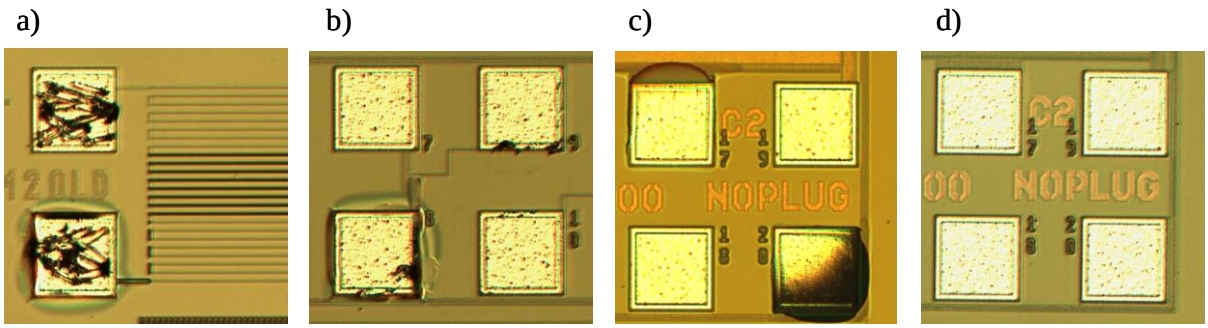


Abbildung 5.5: Mikroskopbilder von Ausschnitten der Wafer, auf denen Stressmigrationstests vorgenommen wurden; a) und b) nach 1400 h bei 350 °C, c) nach 1,5 h bei 450 °C und d) nach 10150 h bei 250 °C; Größe der Pads jeweils 100 µ x 100 µm

Via-Strukturen

Ein ähnliches Bild erhält man auch für die Experimente mit Viaketten. Abbildung 5.6 verdeutlicht dies für jeweils drei Ketten aus 100 Vias zwischen der zweiten und der dritten Metalllage (Via 2 - Ketten) sowie für Ketten aus 100 Vias zwischen der ersten und der dritten Metalllage (Via 1+2 - Ketten). Die Bilder der linken Seite präsentieren auch hier die Ergebnisse der 250 °C-Lagerung, die Bilder der rechten Seite die Resultate der 350 °C-Lagerung.

Bei den Via 2 - Ketten ist nach 10.000 h bei 250 °C keine Widerstandsänderung festzustellen. Bei den Via 1+2 - Ketten hat eine der drei Teststrukturen nach 3000 h einen deutlich höheren Widerstand, was auf Kratzer auf dem Wafer zurückzuführen ist[9]. Bei der 350 °C-Lagerung sind teilweise schon nach 500 h, spätestens aber nach mehr als 2000 h auch hier die Widerstände deutlich erhöht und die zugehörigen Pads oxidiert.

Zusammenfassend lässt sich Stressmigration also weder bei Metallbahnen noch bei den Viaketten nachweisen. Messbare Widerstandszunahmen sind immer auf die Oxidation der Pads zurückzuführen.

[8] Zur Erinnerung: Der Ofen wird zwar mit Stickstoff gespült, ist aber nicht hermetisch dicht.
[9] Für eine bessere Übersichtlichkeit sind auch hier nur die Messwerte vor dem Ausfall dargestellt.

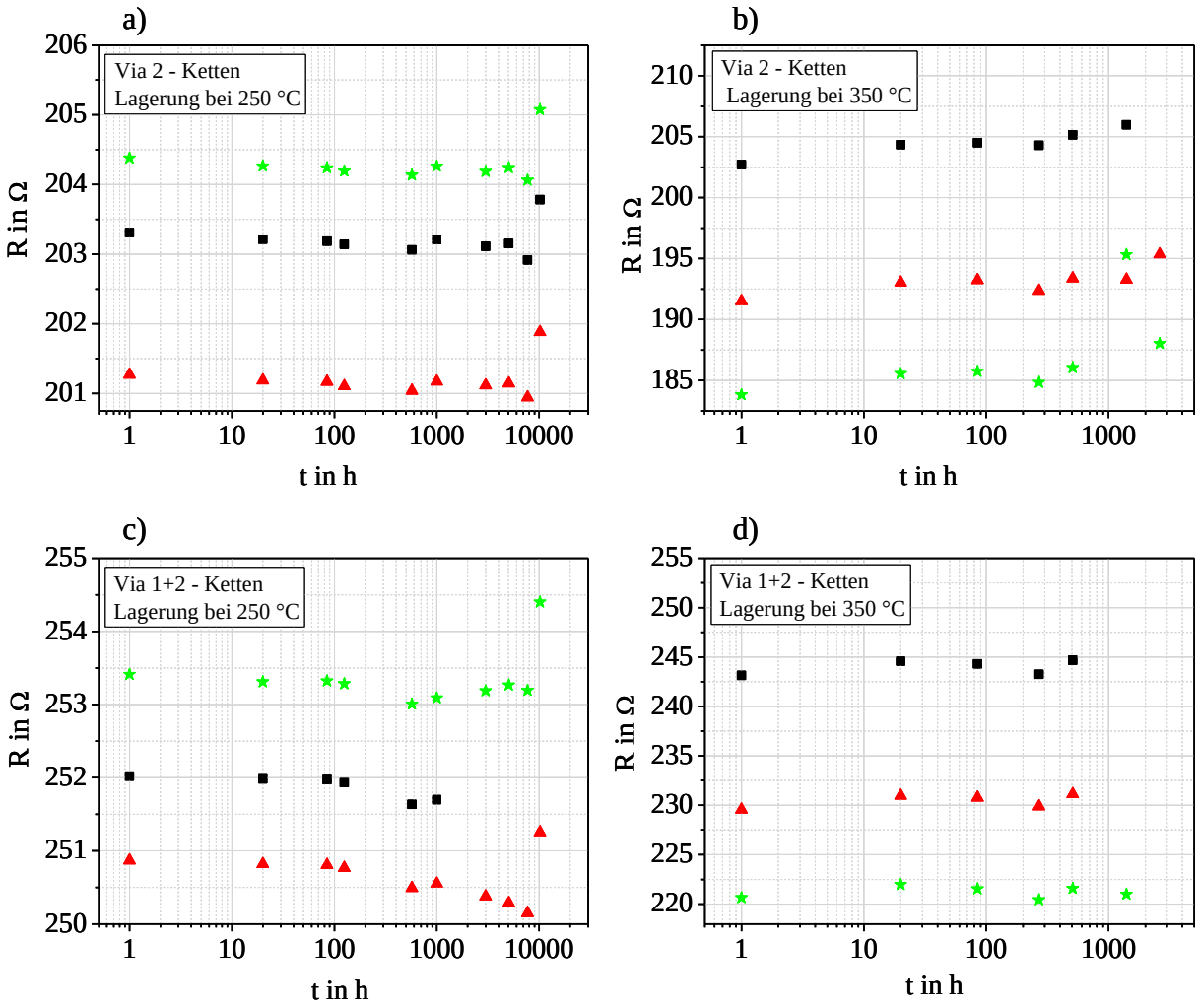


Abbildung 5.6: Widerstände von jeweils drei Viaketten in Abhängigkeit von der Zeit bei einer Lagerungstemperatur von 250 °C (a und c) bzw. 350 °C (b und d); die Werte vor der Lagerung bei $t = 0$ h sind wegen der logarithmischen x-Achsendarstellung bei $t = 1$ h eingezeichnet; je 100 „Via 2“ (Via 2 - Ketten in a) und b)) bzw. 100 „Via 1 und 2 übereinander“ (Via 1+2 - Ketten in c) und d))

Kontakte zu Polysilizium und Aktivgebiet

Im Gegensatz zur Stabilität der Widerstände bei Metallbahnen, Metallmäandern und Viaketten bis zum Zeitpunkt des Ausfalls ergibt sich bei den einzelnen Kontakten und Kontaktketten ein anderes Bild. In Abbildung 5.7 sind auf der linken Seite die Ergebnisse für die Widerstandsänderungen von Einzelkontakten zwischen Metall[10] und Pplus-dotiertem Aktivgebiet (a), zwischen Metall und Nplus-dotiertem Aktivgebiet (c) sowie zwischen Metall und Polysilizium (e) in Abhängigkeit von der Lagerungszeit bei 350 °C dargestellt[11]. Auf der rechten Seite sind die Änderungen der Widerstände von Ketten aus je 36 Kontakten zu sehen[12]. Die Messwerte sind als Mittelwerte mit Standardabweichung angegeben, weil jeweils zwischen fünf und zehn Strukturen vermessen wurden. Alle Strukturen wurden mehr als 1000 h bei 350 °C gelagert, wobei die Poly-Ketten und die Pplus-Ketten nach 1000 h deutlich

[10] Bei „Metall“ handelt sich in diesem Abschnitt immer um die erste Metalllage (Metall 1).
[11] Aus Gründen der Vereinfachung wird im Folgenden für Pplus-dotiertes Aktivgebiet häufig nur „Pplus“, für Nplus-dotiertes Aktivgebiet „Nplus“ und für Polysilizium die Abkürzung „Poly“ verwendet.
[12] Die Widerstände wurden jeweils aus den erhaltenen Spannungswerten bei Raumtemperatur für den angelegten Strombereich von 0,1 mA bis 0,3 mA berechnet.

erhöhte Widerstände zeigten (mindestens 50 %, auf die Pad-Oxidationen zurückzuführen, nicht mehr dargestellt)[13]. Aufgrund der langen Lagerungszeiten (um eine Degradation zu beobachten) wurde von zusätzlichen Tests bei 250 °C abgesehen.

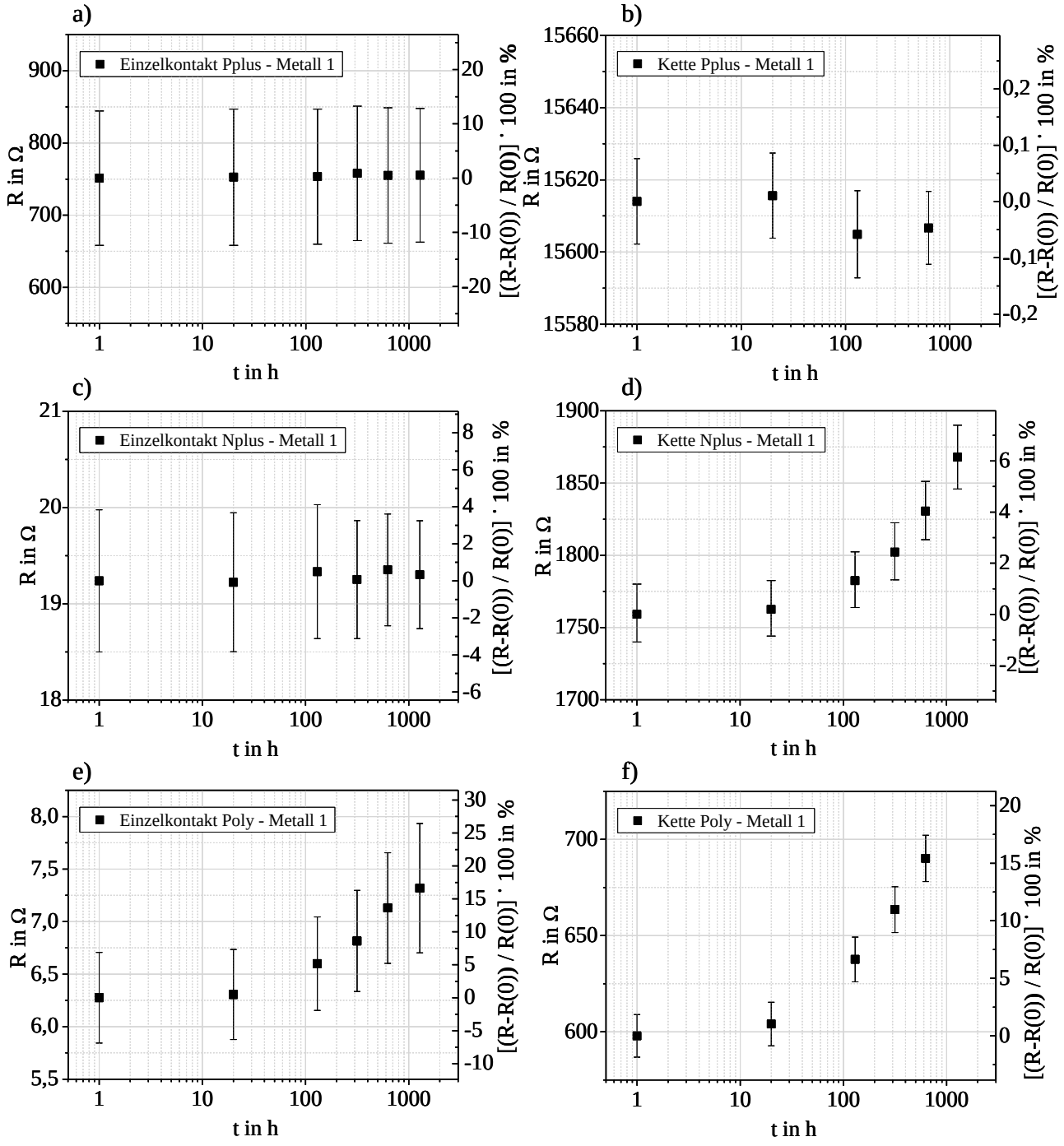


Abbildung 5.7: Mittelwerte der Widerstände von Einzelkontakten (a, c, e) und Kontaktketten (b, d, f) in Abhängigkeit von der Zeit bei einer Lagerungstemperatur von 350 °C; die Werte $R(0)$ vor der Lagerung bei $t = 0$ h sind wegen der logarithmischen x-Achsendarstellung bei $t = 1$ h eingezeichnet; absolute Widerstandswerte auf der linken y-Achse, prozentuale Änderungen im Vergleich zu $t = 0$ h auf der rechten y-Achse; je 36 Kontakte pro Kontaktkette; jeder Mittelwert setzt sich aus 5 (Ketten) bis 10 (Einzelkontakte) Teststrukturen zusammen

[13] Die Messdaten sind wie schon zuvor nur bis zum letzten Wert vor dem Ausfall dargestellt.

Die Widerstände der einzelnen Aktivgebietskontakte bleiben während der gesamten Lagerungszeit stabil (Abbildung 5.7 a) und c)), während sich die Widerstände der einzelnen Metall-Polysilizium-Kontakte mit der Zeit erhöhen (Abbildung 5.7 e)). Auch bei den Kontaktketten gibt es Unterschiede. Die Widerstände der Pplus-Metall-Ketten zeigen bis zum Ausfall nach mehr als 1000 h keine merklichen Änderungen (Abbildung 5.7 b)). Die anderen beiden Ketten (Abbildung 5.7 d) und f)) weisen hingegen ähnliche Widerstandsanstiege wie die Polysilizium-Metall-Einzelkontakte auf, wobei die Änderungen der einzelnen Polysilizium-Metall-Kontaktwiderstände und der entsprechenden Ketten größer sind als die der Nplus-Metall-Ketten.

Der in den Abbildungen d), e) und f) in einem Zeitraum von 1000 h erkennbare Widerstandsanstieg war in dieser Form bei den Metallbahnen und Vias nicht zu sehen.

Die Kontaktketten bestehen aus 18 Abschnitten, bei denen immer zwei Kontakte auf einem Aktivgebiets- bzw. Polysiliziumsteg miteinander verbunden sind. Da der Widerstand dieses Stegstücks aber wesentlich größer ist als die Kontaktwiderstände, zeigen die Bilder der rechten Seite in Abbildung 5.7 gar nicht die tatsächliche Degradation der Kontaktwiderstände in der Kette sondern die Degradation der Widerstände von Nplus bzw. Polysilizium selbst. In Abbildung 5.8 sind deshalb die Ergebnisse der Lagerungen von separaten Pplus-, Nplus- und Poly-Widerstandsstrukturen (als Square-Widerstände) bei 250 °C und 350 °C dargestellt. Auch hier ist zu erkennen, dass innerhalb von 1000 h bei 350 °C die Widerstände von Nplus und Polysilizium ansteigen während die Widerstände von Pplus stabil bleiben. Bei 250 °C ist innerhalb des Beobachtungszeitraumes keine Änderung zu erkennen.

Der Grund für den Anstieg des hochohmigen, mit Phosphor n-dotierten Polysiliziumwiderstands ist nach Rydberg *et al.* das Aufbrechen von schwachen Wasserstoffbindungen in bzw. an den Korngrenzen durch Temperaturstress [RYD00]. Die aufgebrochenen Bindungen bilden so genannte „dangling bonds“, die für den Elektronenfluss durch das Polysilizium als Störstellen wirken und so den Widerstand erhöhen. Phosphor selbst kann den Wasserstoff wieder binden und dadurch verhindern, dass der Wasserstoff wieder in den Korngrenzen gebunden wird. Wo genau der Wasserstoff seinen Ursprung hat (BPSG-Layer, Nitrid-Passivierung, …) ist auch in der Literatur noch in der Diskussion.

Die gleiche Erklärung gilt prinzipiell auch für die Widerstandserhöhung der durch Phosphor n-dotierten Aktivgebietsstege. Die beobachtete Änderung der gemessenen Widerstände ist aber aus zwei Gründen geringer als bei Polysilizium. Zum einen ist die Dotierung der Nplus-Gebiete mit $8 \cdot 10^{19}$ 1/cm³ geringer als bei Polysilizium, wo die $POCl_3$-Belegung (Phosphoroxychlorid) zu einem deutlich höheren Wert führt, zum andern handelt es sich bei den dotierten Aktivgebieten um einkristallines Silizium. Es gibt also keine Korngrenzen, an denen Wasserstoffbindungen aufgebrochen werden können. „dangling bonds“ könnten aber an der Siliziumgrenzfläche entstehen und dadurch die Beweglichkeit im Siliziumfilm reduzieren.

Bei den mit Bor p-dotierten Aktivgebietsbereichen ist kein Phosphor vorhanden, das verhindern könnte, dass die an den Korngrenzen aufgebrochenen Wasserstoffbindungen dort wieder abgesättigt werden. Rydberg *et al*. haben an mit Bor dotiertem Polysilizium gezeigt, dass dort Fluor die erneute Bindung von aufgebrochenen Wasserstoffbindungen verhindert [RYD01]. Da bei in dem hier vorliegenden Fall aber kein Fluor verwendet wurde und zudem auch keine Korngrenzen vorhanden sind, gibt es nur wenige oder keine aufgebrochenen Bindungen. Zumindest im Zeitraum von 1000 h bei 350 °C ist deshalb an den Pplus-Widerständen keine Degradation zu beobachten.

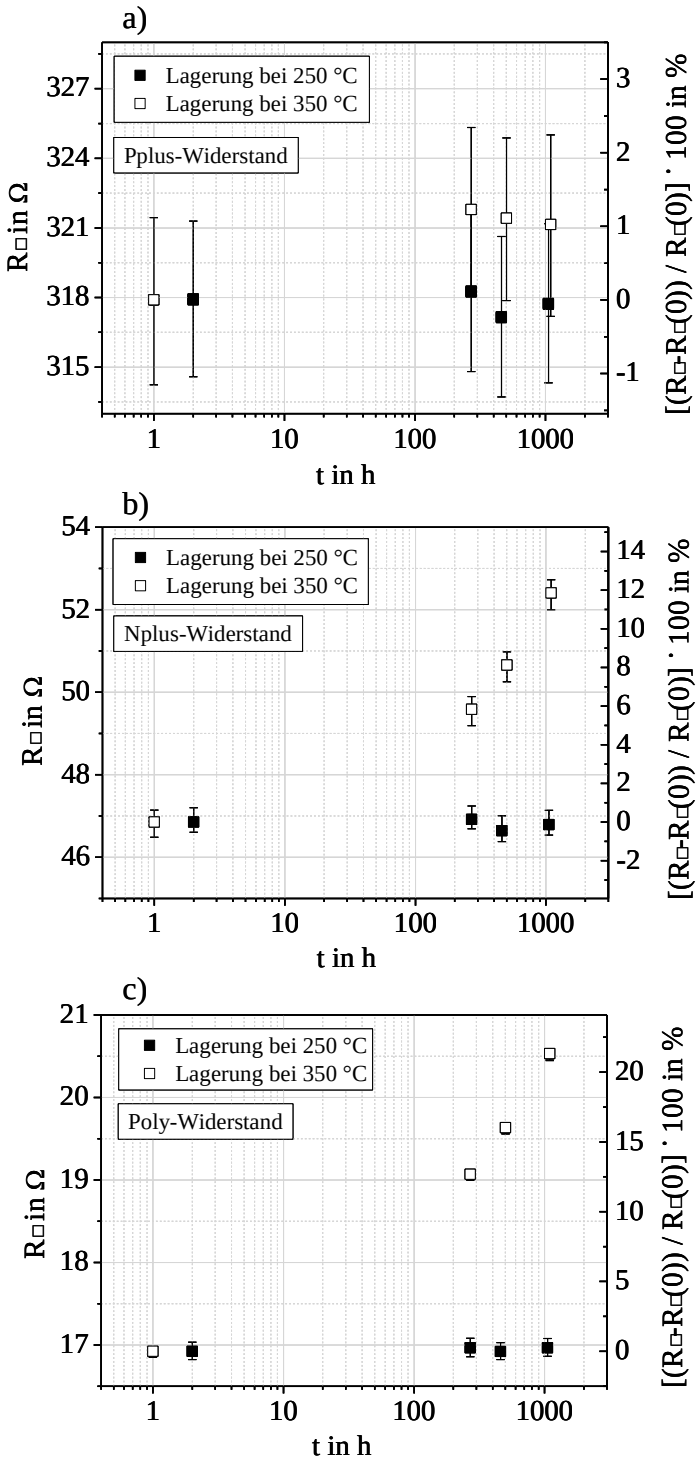


Abbildung 5.8: Mittelwerte der Square-Widerstände $R_{\square} = R_{gemessen} \cdot$ Bahnbreite / Bahnlänge von Pplus- (a), Nplus- (b) und Polysilizium- (c) Bahnen in Abhängigkeit von der Zeit bei einer Lagerungstemperatur von 250 °C (ausgefüllte Kästchen) bzw. 350 °C (leere Kästchen); die Werte $R_{\square}$ (0) vor der Lagerung bei t = 0 h sind wegen der logarithmischen x-Achsendarstellung für die 350 °C-Lagerung bei t = 1 h und für die 250 °C-Lagerung bei t = 2 h eingezeichnet; absolute Widerstandswerte auf der linken y-Achse, prozentuale Änderungen im Vergleich zu den 350 °C-Mittelwerten bei t = 0 h auf der rechten y-Achse; Bahnlänge l = 110 µm, Bahnbreite b = 20 µm; jeder Mittelwert setzt sich aus 10 Teststrukturen zusammen

Um wieder auf die Kontakte zurückzukommen (siehe Abbildung 5.7), scheinen an den Aktivgebiets-Kontakten offene Wasserstoffbindungen keine Rolle zu spielen, da sich der

Widerstand in der gemessenen Zeit nicht ändert. Bei den Polysilizium-Metall-Kontakten (Abbildung 5.7 e)) hingegen ist die Widerstandszunahme auf die „dangling bonds“ im Polysilizium zurückzuführen. Der Wolframkontakt selbst degradiert nicht.

Dass das Aufbrechen der Wasserstoffbindungen tatsächlich der Grund für die Widerstandszunahmen ist, kann auch durch die Übereinstimmung der hier vorliegenden Ergebnisse mit dem von Rydberg *et al.* beschriebenen Zusammenhang zwischen den Widerständen und der Zeit bestätigt werden [RYD00]. Rydberg *et al.* geben die zeitliche Entwicklung eines Widerstands $R(t)$ folgendermaßen an:

$$\ln\left(\frac{R(t)}{R(0)}\right) = a \cdot \left[1 - \exp\left(-b \cdot t\right)\right] \qquad [5.4]$$

Dabei ist $R(0)$ der Widerstand vor der Lagerung (t = 0 h) und a und b sind temperaturabhängige Konstanten. Die Gültigkeit dieser Gleichung für die Diagramme in Abbildung 5.7 d), e) und f) kann mit Abbildung 5.9 bestätigt werden. Dort sind die Logarithmen der Quotienten aus den Widerständen $R(t)$ zum Zeitpunkt t und den Widerständen $R(0)$ in Abhängigkeit von der Zeit dargestellt und Fitkurven in die Messpunkte gelegt. Die Diagramme in Abbildung 5.9 a) und b), c) und d) sowie e) und f) beinhalten die gleichen Daten. Auf der linken Seite ist die x-Achse aber zum Vergleich mit Abbildung 5.7 logarithmisch, auf der rechten Seite ist sie linear, um auch den Wert bei t = 0 h mit in den Fit einbeziehen zu können. In allen drei Fällen spiegeln die Fitkurven Gleichung 5.4 sehr gut wider.

Nach [RYD00] sind zeitliche Veränderungen des Polysilizium-Widerstands bei kombiniertem Temperatur- und Stromstress vor allem auf den Einfluss der Temperatur und weniger auf den des Stromes zurückzuführen. Die aufbrechenden Wasserstoffbindungen könnten deshalb auch der Grund sein, weshalb beim Polysilizium-Metall-Kontakt in Abbildung 5.3 eine leichte Widerstandszunahme mit der Zeit beobachtet werden konnte.

Fazit

In den Untersuchungen konnten während 10.000 h bei 250 °C und etwa 2000 h bei 350 °C keine Stressmigration an Metallbahnen, Metallmäandern und Viaketten festgestellt werden. Abrupte Widerstandsanstiege sind immer auf die Oxidation der Pads zurückzuführen. Dafür treten aber Änderungen von Kontakt- und (Poly-)siliziumwiderständen durch Temperaturlagerung als Folge aufgebrochener Wasserstoffbindungen an den Korngrenzen auf. Vor allem Poly-Widerstände und Poly-Metall-Kontakte sind davon betroffen. Die beobachteten Widerstandsänderungen liegen aber nach etwa 1000 h bei 350 °C bei weniger als 20 % und sind damit im Verhältnis zu anderen Zuverlässigkeitsproblemen bei derselben Temperatur (siehe andere Kapitel) eher unkritisch.

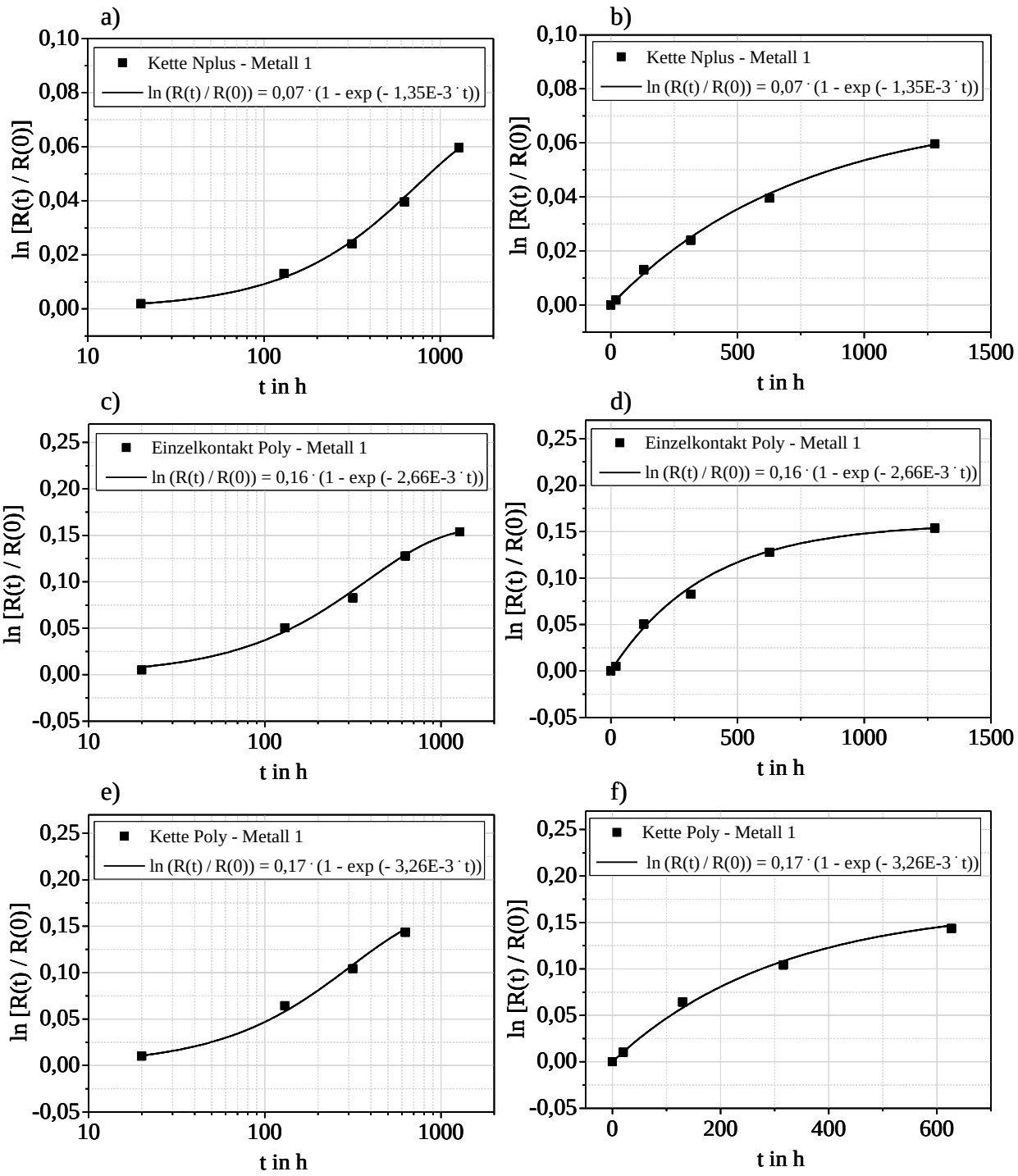


Abbildung 5.9: Logarithmus des Quotienten aus $R(t)$ und $R(0)$ mit $R(t)$ als Mittelwert der jeweiligen Widerstände zum Zeitpunkt t und $R(0)$ als Mittelwert der jeweiligen Widerstände zum Zeitpunkt t = 0 h in Abhängigkeit von der Zeit bei einer Lagerungstemperatur von 350 °C; Widerstände $R(t)$ und $R(0)$ aus den Diagrammen d, e, und f aus Abbildung 5.7; Fitkurven nach Gleichung 5.4; logarithmische x-Achsendarstellung (a, c, e) und lineare x-Achsendarstellung (b, d, f); bei der logarithmischen Darstellung sind die Werte bei t = 0 h nicht eingezeichnet, um den Fit nicht zu verfälschen; die Diagramme a, c, und e entsprechen den Diagrammen d, e, und f aus Abbildung 5.7; je 36 Kontakte pro Kontaktkette; jeder Mittelwert setzt sich aus 5 (Ketten) bis 10 (Einzelkontakte) Teststrukturen zusammen; die aufgrund der Oxidation der Pads deutlich erhöhten Widerstandswerte nach 1000 h Lagerungszeit wurden nicht mit einbezogen

Kapitel 6

Charakterisierung und Langzeituntersuchungen von Transistoren und einfachen Grundschaltungen

In den vorangegangenen Kapiteln wurde die Zuverlässigkeit einzelner Komponenten einer Technologie, wie dem Gateoxid und der Metallisierung, analysiert. Dieses Kapitel behandelt nun die Untersuchung der Zuverlässigkeit von Transistoren und einfachen Testschaltungen. Dazu wurden Transistoren, Ringoszillatoren und Bandgap-Referenzen bis zu einer Temperatur von 450 °C charakterisiert und ihre Stabilität bei 250 °C und, wenn möglich, auch bei 350 °C über einen längeren Zeitraum beobachtet. Eine weitere Zuverlässigkeitsanalyse stellt die Untersuchung von zwei Effekten dar, die die Funktionalität von Transistoren und damit auch von Schaltungen einschränken können: der Einfluss von „heißen" Ladungsträgern (engl. „Hot Carrier" oder HCI) und die Parameterinstabilität bei negativer Gate-Source-Spannung (engl. „Negative Bias Temperature Instability" oder NBTI).

6.1 MOSFET-Transistoren

6.1.1 MOSFET-Transistoren bei hohen Umgebungstemperaturen

Basis der hochtemperaturtauglichen H10-Technologie ist ein CMOS-Prozess auf Dünnfilm-Silicon-on-Insulator (SOI). Der Begriff CMOS steht für Complementary Metal Oxide Semiconductor und bezeichnet die gemeinsame Realisierung zweier MOSFET-Transistoren (Metal Oxide Semiconductor Field Effect Transistors), eines n-Kanal-MOSFETs (kurz: NMOS) und eines p-Kanal-MOSFETs (kurz: PMOS) auf demselben Substrat. Im Falle einer SOI-Technologie sind beide Transistoren im selben Film realisiert und durch einen Feldoxidbereich voneinander getrennt (siehe auch Abbildung 2.1 in Kapitel 2). Schaltungstechnisch können ein NMOS- und ein PMOS-Transistor zusammen so aufgebaut werden, dass sie als Inverter wirken. Anwendung findet dieser beispielsweise im Ringoszillator (siehe Abschnitt 6.2). Die wesentlichen Vorteile der CMOS-Technologie sind der geringe Stromverbrauch und generell die Realisierung digitaler Schaltungen.

Um die Auswirkungen von hohen Umgebungstemperaturen auf die Funktion einer Schaltung zu untersuchen, müssen zunächst die temperaturbedingten Veränderungen der einzelnen Bauteile untersucht werden. Dabei spielen vor allem die Eigenschaften der verwendeten Materialien, hier des Siliziums, eine Rolle. Zwei wesentliche Punkte sind in diesem Zusammenhang zu nennen, die intrinsische Ladungsträgerkonzentration n_i und die Ladungsträgerbeweglichkeit μ.

Die intrinsische Ladungsträgerkonzentration n_i

Die intrinsische Ladungsträgerkonzentration n_i steigt mit der Temperatur T an [MUL86], [COL97]:

$$n_i = 3{,}9 \cdot 10^{16} \cdot T^{3/2} \cdot \exp\left(-\frac{E_g}{2 \cdot k \cdot T}\right) \qquad \text{(Einheit: cm}^{-3}\text{)} \qquad [6.1]$$

Dabei ist k die Boltzmann-Konstante und E_g die Bandlücke von Silizium. Letztere ist im Bereich von Raumtemperatur bis 250 °C nahezu temperaturunabhängig [COL97] und nimmt mit höherer Temperatur leicht ab [MUL86], was aber vernachlässigt werden kann [WER96].

Die Beweglichkeit μ

Die Beweglichkeit μ von Ladungsträgern wird durch verschiedene Streuprozesse, wie der Gitterstreuung (Phononenstreuung) und der Streuung an Defekten oder Störstellen, bestimmt. Die Zeit zwischen zwei aufeinander folgenden Streuprozessen nimmt mit der Temperatur ab, wodurch die Beweglichkeit reduziert wird. Der Zusammenhang zwischen Beweglichkeit μ und Temperatur T ist laut [GIR65]:

$$\mu = \mu_0 \cdot \left(\frac{T}{T_0}\right)^n \qquad [6.2]$$

Die Beweglichkeit μ_0 ist hierbei die Beweglichkeit bei der Temperatur T_0. Der Exponent n wird in den meisten Fällen mit -3/2 angegeben, kann aber mit dem Material, der Dotierung oder auch bei sehr hohen Temperaturen (400 °C) davon abweichen [REG02]. Die Beweglichkeiten für Elektronen und Löcher, μ_n und μ_p, unterscheiden sich um einen Faktor von ungefähr drei:

$$\mu_n \approx 3 \cdot \mu_p \qquad [6.3]$$

Die Temperaturabhängigkeit von intrinsischer Ladungsträgerkonzentration n_i und Beweglichkeit μ haben Einfluss auf die Bauteilparameter, bei Transistoren vor allem auf die Schwellenspannung, die Leckströme und den Sättigungsstrom.

Die Schwellenspannung U_{th}

Für die Schwellenspannung U_{th} eines MOSFET-Transistors in SOI-Technologie gilt [COL97]:

$$U_{th} = \Phi_{MS} + 2 \cdot \Phi_F - \frac{Q_{ox}}{C_{ox}} + \frac{Q_{depl}}{C_{ox}} \qquad [6.4]$$

Dabei ist Φ_{MS} die Differenz der Austrittsarbeiten zwischen Gate- und Filmmaterial, Φ_F das Fermipotential, Q_{ox} die Ladung im Gateoxid, Q_{depl} die Ladung im Verarmungsbereich und C_{ox} die Kapazität des Gateoxids[1].

[1] Die angegebenen Ladungen und Kapazitäten sind korrekterweise Ladungen und Kapazitäten *pro Fläche*. Der Einfachheit und Konsistenz halber wird aber nur von Ladungen und Kapazitäten gesprochen.

Weiterhin gilt:

$$\Phi_{MS} = -\frac{E_g}{2} - \Phi_F \quad \text{mit} \quad \Phi_F = \frac{k \cdot T}{q} \cdot \ln\left(\frac{N_{a,d}}{n_i}\right) \qquad [6.5] \text{ und } [6.6]$$

Für Transistoren mit einem nur teilweise verarmten Siliziumfilm (engl. "partially depleted") wird Q_{depl} durch die maximale Tiefe der Verarmungsschicht x_{dmax} bestimmt (Gleichungen 6.7 und 6.8). Für vollständige Verarmung (engl. „fully depleted") hängt Q_{depl} von der Dicke des Siliziumfilms d_{si} ab (Gleichung 6.9).

$$Q_{depl} = q \cdot N_{a,d} \cdot x_{dmax} \quad \text{mit} \quad x_{dmax} = \sqrt{\frac{4 \cdot \varepsilon_{si} \cdot \Phi_F}{q \cdot N_{a,d}}} \qquad [6.7] \text{ und } [6.8]$$

$$Q_{depl} = q \cdot N_{a,d} \cdot \frac{d_{si}}{n} \qquad [6.9]$$

E_g ist dabei die Bandlücke von Silizium, k die Boltzmann-Konstante, T die Temperatur, q die Ladung eines Elektrons bzw. Lochs, $N_{a,d}$ die Kanaldotierung bei NMOS bzw. PMOS, n_i die intrinsische Ladungsträgerkonzentration, ε_{Si} die Dielektrizitätskonstante von Silizium und n ein Wert zwischen eins und zwei, der sich aus dem Einfluss des Back-Gates ergibt.

Für die Temperaturabhängigkeit der Schwellespannung von nur teilweise verarmten Dünnfilm-SOI-Transistoren liegt derselbe Zusammenhang wie für Bulk- oder Dickfilm-SOI-Transistoren [GRO90], [COL97] vor:

$$\frac{dU_{th}}{dT} = \frac{d\Phi_F}{dT} \cdot \left(1 + \frac{q}{C_{ox}} \sqrt{\frac{\varepsilon_{si} \cdot N_{a,d}}{k \cdot T \cdot \ln(N_{a,d}/n_i)}}\right) \qquad \text{(Einheit: V/K)} \qquad [6.10]$$

Dabei gilt für die Temperaturabhängigkeit des Fermipotentials Φ_F:

$$\frac{d\Phi_F}{dT} = 8{,}63 \cdot 10^{-5} \cdot \left(\ln(N_{a,d}) - 38{,}2 - \frac{3}{2} \cdot [1 + \ln(T)]\right) \qquad \text{(Einheit: V/K)} \qquad [6.11]$$

Für vollständig verarmte Dünnfilm-SOI-Transistoren reduziert sich Gleichung 6.10 auf den ersten Term, wenn in erster Näherung eine Temperaturunabhängigkeit für n angenommen wird [GRO90], [COL97]. Es gilt dann:

$$\frac{dU_{th}}{dT} = \frac{d\Phi_F}{dT} \qquad \text{(Einheit: V/K)} \qquad [6.12]$$

Die Schwellenspannung hängt damit über das Fermipotential Φ_F und die intrinsische Ladungsträgerkonzentration n_i von der Temperatur ab.

Für die Änderung der Schwellenspannung mit der Temperatur ist bei Bulk-Transistoren von 25 °C bis zu einer Temperatur von 300 °C laut Shoucair *et al*. eine Näherung durch ein Polynom erster Ordnung als Lösung (von Gleichung 6.10) ausreichend [SHO84][2]:

$$\left|U_{\mathrm{th}}\right| = \left|U_{\mathrm{th,0}}\right| - a \cdot T \qquad [6.13]$$

Hierbei bezeichnet $U_{\mathrm{th,0}}$ die Schwellenspannung bei Raumtemperatur und a eine positive Konstante. Auch für vollständig verarmte Dünnfilm-SOI-Transistoren gilt ein linearer Zusammenhang zwischen der Schwellenspannung und der Temperatur bis etwa 250 °C [GRO90], [VAN08]. Der Betrag der Schwellenspannung sinkt also linear mit steigender Temperatur, wobei die Änderung bei vollständig verarmten Dünnfilm-SOI-Transistoren um einen Faktor von etwa zwei bis drei kleiner ist als bei Bulk- oder Dickfilm-SOI-Transistoren [GRO90]. Ob die Näherung eines linearen Zusammenhangs zwischen Schwellenspannung und Temperatur auch noch für höhere Temperaturen gilt, wird in Abschnitt 6.1.2 untersucht.

Der Leckstrom I_{leck}

Der gesamte Leckstrom I_{leck}, der im Off-Zustand eines Transistors fließt, setzt sich aus mehreren Komponenten zusammen. Bei der Betrachtung der Temperaturabhängigkeit spielen vor allem der pn-Leckstrom I_{pn} der in Sperrrichtung gepolten pn-Übergänge (bei $U_{\mathrm{GS}} \leq 0$ V) und der Subtreshold-Leckstrom I_{sub} (für $U_{\mathrm{GS}} < U_{\mathrm{th}}$) eine Rolle. Weitere Leckströme, vor allem bei kurzen Kanallängen, können bei stark negativen (NMOS) bzw. stark positiven (PMOS) Gate-Source-Spannungen (engl. Gate Induced Drain Leackage oder GIDL) oder bei hohen Drain-Source-Spannungen durch die Drain-induzierte Potentialbarrien-Verringerung (engl. Drain Induced Barrier Lowering oder DIBL) entstehen.

Der pn-Leckstrom (oder np-Leckstrom für in Sperrrichtung gepolte np-Übergänge) kann mit folgender Gleichung beschrieben werden [SZE81], [COL97]:

$$I_{\mathrm{pn,np}} = q \cdot A \cdot \left(\frac{D_{\mathrm{p,n}}}{\tau_{\mathrm{p,n}}} \right) \cdot \frac{n_{\mathrm{i}}^2}{N_{\mathrm{d,a}}} + q \cdot A \cdot \frac{n_{\mathrm{i}} \cdot W}{\tau_{\mathrm{p,n}}} \qquad [6.14]$$

Mit q wird dabei die Ladung eines Elektrons bzw. Lochs bezeichnet, A ist die Fläche des pn-Übergangs, D_{n} bzw. D_{p} der Diffusionskoeffizient für Elektronen bzw. Löcher, τ_{p} bzw. τ_{n} die Lebensdauer eines Lochs bzw. Elektrons, n_{i} die intrinsische Ladungsträgerkonzentration, N_{a} bzw. N_{d} die Kanaldotierung und W die Breite der Raumladungszone.

Aus den Gleichungen 6.14 und 6.1 lässt sich schlussfolgern, dass der pn-Leckstrom mit steigender Temperatur zunimmt. Der erste Term in Gleichung 6.14 stellt den Diffusionsstrom dar. Er ist proportional zu n_{i}^2, entfällt aber bei vollständiger Verarmung. Die zweite Komponente ergibt sich aus dem Generations-Rekombinations-Strom. Dieser ist proportional zu n_{i} und dominiert, solange vollständige Verarmung vorliegt.

Die Fläche A des pn-Übergangs und das Produkt $A \cdot W$ als das Volumen der Raumladungszone sind bei SOI-Transistoren wesentlich kleiner als bei Transistoren auf Bulk-Wafern. Deshalb

[2] Für diese Näherung wird nur der erste Term in Gleichung 6.10 berücksichtigt, d. h. im Prinzip Gleichung 6.12.

sind die Leckströme bei SOI-Transistoren deutlich geringer und SOI ist als Basismaterial für Hochtemperaturanwendungen gut geeignet.

Für den Subtreshold-Leckstrom gilt [ROY03]:

$$I_{\text{sub}} = \left[\mu_{\text{n,p}} \cdot C_{\text{ox}} \cdot \frac{w}{l} \cdot (n-1) \cdot \left(\frac{k \cdot T}{q} \right)^2 \cdot \exp\left(q \cdot \frac{U_{\text{GS}} - U_{\text{th}}}{n \cdot k \cdot T} \right) \cdot \left(1 - \exp\left(-\frac{q \cdot U_{\text{DS}}}{k \cdot T} \right) \right) \right] \qquad [6.15]$$

μ ist die Beweglichkeit von Elektronen bzw. Löchern, C_{ox} die Gateoxidkapazität, w die Kanalweite und l die Kanallänge des Transistors, k ist die Boltzmann-Konstante, T die Temperatur, U_{GS} die Gate-Source-Spannung, U_{th} die Schwellenspannung, U_{DS} die Drain-Source-Spannung und n_{B} der Subtreshold-Swing-Koeffizient oder Body-Effekt-Koeffizient. Für einen nur teilweise verarmten Siliziumfilm gilt für den Koeffizienten n_{B} [COL97]:

$$n_{\text{B}} \approx \left(1 + \frac{C_{\text{depl}}}{C_{\text{ox}}} \right) \approx \left(1 + \frac{\varepsilon_{\text{si}}}{x_{\text{dmax}} \cdot C_{\text{ox}}} \right) \qquad [6.16]$$

Dabei ist C_{depl} die Depletionkapazität, ε_{Si} die Dielektrizitätskonstante von Silizium und x_{dmax} die maximale Tiefe der Verarmungsschicht. Im Fall vollständiger Verarmung gilt $x_{\text{dmax}} = d_{\text{si}}$, wobei d_{si} die Dicke des Siliziumsfilms ist und für eine exakte Lösung noch die Back-Gate-Kapazität berücksichtigt werden müsste. Der Subtreshold-Leckstrom hängt also auch von der Temperatur ab. Vor allem für hohe Temperaturen kann er über dem pn-Leckstrom liegen und damit eine Rolle spielen [WER96].

Der Sättigungsstrom $I_{\text{sätt}}$

Der Sättigungsstrom $I_{\text{sätt}}$ für in Flussrichtung gepolte pn-Übergänge nimmt mit der Temperatur ab. Der Grund dafür ist die mit steigender Temperatur kleiner werdende Beweglichkeit der Ladungsträger. Auf der anderen Seite spielt aber auch die Temperaturabhängigkeit der Schwellenspannung eine Rolle.

$$I_{\text{sätt}} = \frac{1}{2} \cdot \mu_{\text{n,p}} \cdot C_{\text{ox}} \cdot \frac{w}{l} \cdot \left(U_{\text{GS}} - U_{\text{th}} \right)^2 \quad \text{für} \quad |U_{\text{DS}}| \geq |U_{\text{GS}} - U_{\text{th}}| \qquad [6.17]$$

Dabei sind μ_{n} und μ_{p} die Ladungsträgerbeweglichkeiten von Elektronen und Löchern, C_{ox} ist die Kapazität des Gateoxids, w die Kanalweite und l die Kanallänge des Transistors, U_{GS} die Gate-Source-Spannung, U_{th} die Schwellenspannung und U_{DS} die Drain-Source-Spannung.

Der Einfluss der Temperatur auf die Bauelemente hat, wie bereits angedeutet, auch Auswirkung auf die Funktion einer Schaltung. Die Abnahme der Schwellenspannung kann sowohl analoge als auch digitale Schaltungen beeinflussen. Da der Leckstrom mit der Temperatur zunimmt, steigt auch die Ruhestromaufnahme einer Schaltung, andererseits nimmt die Stromtreibefähigkeit der Transistoren ab und die Schaltung wird langsamer. Darauf wird in den Abschnitten 6.2 und 6.3 eingegangen.

6.1.2 Charakterisierung von PMOS- und NMOS-Transistoren

Abbildung 6.1 zeigt Eingangskennlinien, d.h. den Drain-Source-Strom I_{DS} in Abhängigkeit von der Gate-Source-Spannung U_{GS} eines NMOS-Transistors (Abbildung 6.1 a) und b)) und eines PMOS-Transistors (Abbildung 6.1 c) und d)) für Temperaturen zwischen Raumtemperatur (28 °C) und 450 °C.

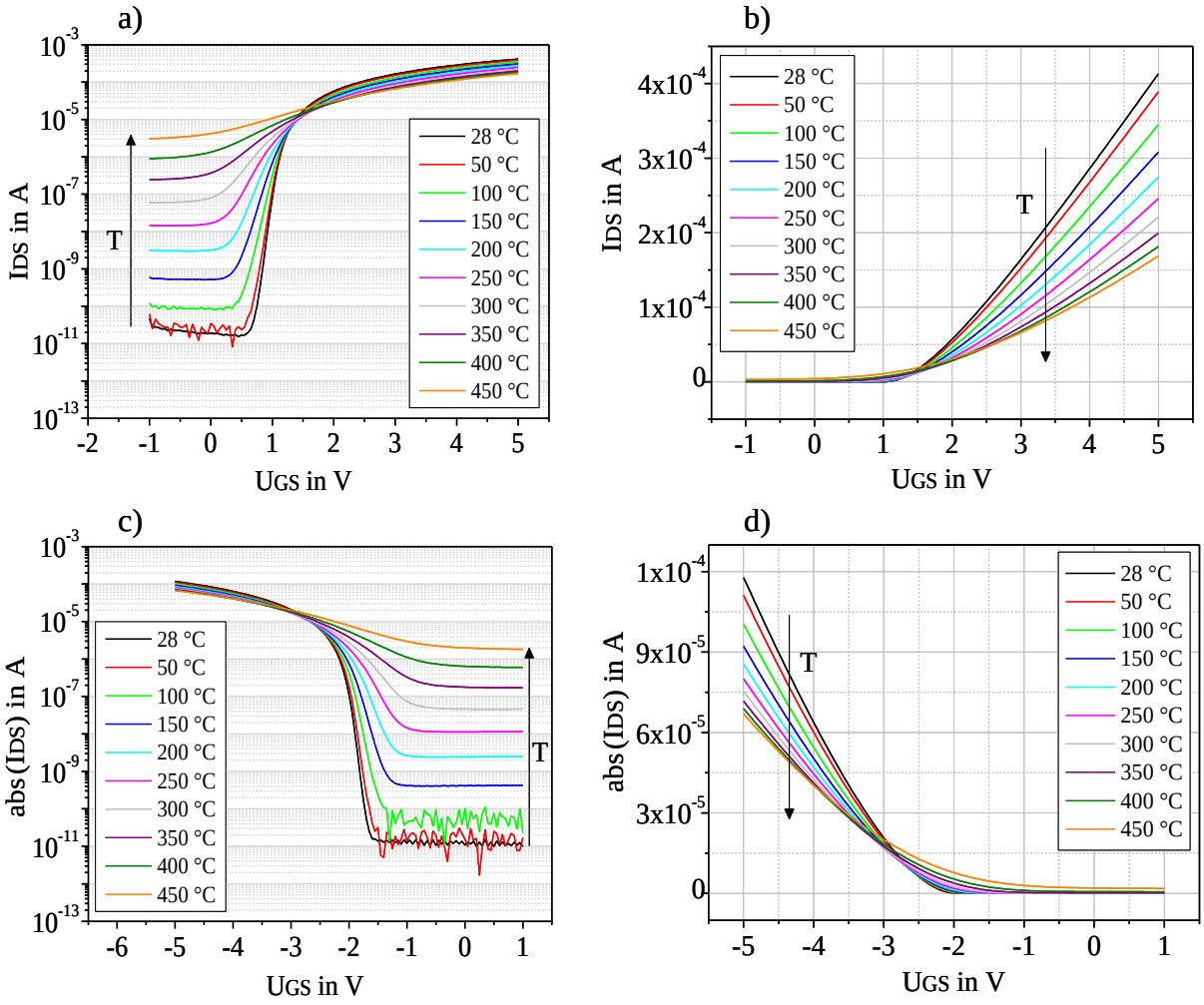


Abbildung 6.1: Eingangskennlinien von NMOS- und PMOS-Transistoren mit einer Kanallänge l = 1,6 µm und einer Kanalweite w = 3,6 µm für Temperaturen zwischen Raumtemperatur und 450 °C; U_{DS} = 5 V (NMOS) sowie U_{DS} = -5 V (PMOS), U_S = 0 V und U_{BG} = 0 V; a) NMOS mit logarithmischer Auftragung der *y*-Achse; b) NMOS mit linearer Auftragung der *y*-Achse; c) PMOS mit logarithmischer Auftragung der *y*-Achse; d) PMOS mit linearer Auftragung der *y*-Achse

Beide Transistoren haben eine Kanallänge von l = 1,6 µm und eine Kanalweite von w = 3,6 µm. Die Messungen wurden an Transistoren im Gehäuse im Ofen durchgeführt. In den Diagrammen a) und c) auf der linken Seite ist die Auftragung der *y*-Achse logarithmisch, in den Diagrammen b) und d) auf der rechten Seite ist sie linear. Für die PMOS-Transistoren ist jeweils der Betrag des Drain-Source-Stromes aufgetragen, damit eine logarithmische Darstellung möglich ist. In den Abbildungen ist U_{DS} = ± 5 V. Für die Berechnung der Schwellenspannung wurden noch weitere Eingangskennlinien bei einer Drain-Source-Spannung von ± 0,1 V aufgenommen. Bei allen Messungen lag das Sourcepotential auf 0 V

und das Drainpotential war bei NMOS und PMOS symmetrisch dazu, um beide Transistoren besser vergleichen zu können. Das Back-Gate lag in allen Fällen auf einem Potential von 0 V.

In den Diagrammen a) und c) von Abbildung 6.1 ist die Zunahme des Leckstromes mit der Temperatur sowohl für den NMOS- als auch für den PMOS-Transistor klar zu erkennen. Die Diagramme b) und d) von Abbildung 6.1 verdeutlichen das Verhalten der Transistoren im Sättigungszustand, wo der Drain-Source-Strom mit steigender Temperatur sichtbar abnimmt. Bevor beide Strombereiche genauer betrachtet werden, wird zunächst die Änderung der Schwellenspannung mit der Temperatur analysiert.

Die Schwellenspannung U_{th}

Abbildung 6.2 zeigt die Schwellenspannungen von NMOS und PMOS in Abhängigkeit von der Temperatur. Die Schwellenspannungen wurden über die Tangentenmethode berechnet[3]. Bei Raumtemperatur beträgt die Schwellenspannung des NMOS etwa 1 V und die des PMOS etwa -2 V.

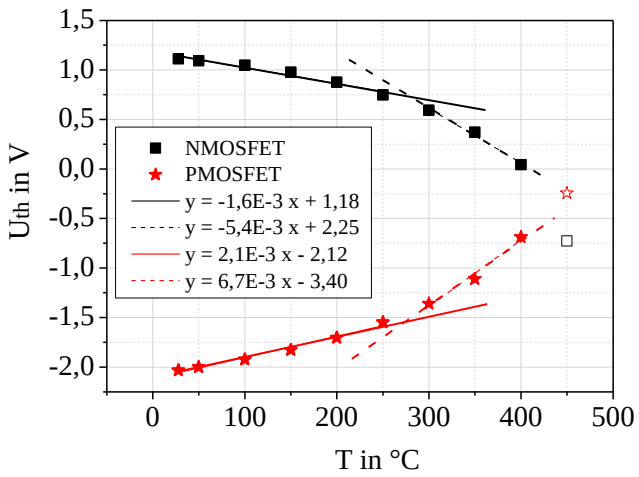


Abbildung 6.2: Schwellenspannungen von NMOS- (Kästchen) und PMOS-Transistoren (Sterne) für Temperaturen zwischen Raumtemperatur und 400 °C; bei 450 °C als leere Symbole; berechnet über die Tangentenmethode aus Eingangskennlinien $I_{DS}(U_{GS})$ mit U_{DS} = 0,1 V (NMOS) bzw. U_{DS} = -0,1 V (PMOS); U_S = 0 V und U_{BG} = 0 V

Der Betrag der Schwellenspannung der beiden Transistoren nimmt, wie in Abschnitt 6.1.1 erläutert, mit der Temperatur ab. Dabei verläuft die Schwellenspannungsänderung mit der Temperatur bei NMOS und PMOS fast symmetrisch. Die Abnahme ist in erster Näherung jeweils im Bereich zwischen Raumtemperatur und etwa 275 °C linear. Für Temperaturen oberhalb von 275 °C nimmt die Schwellenspannung deutlich stärker ab, ist aber auch zwischen 275 °C und 400 °C annähernd linear. Der Unterschied der Steigungen der Fit-Geraden in den beiden Temperaturbereichen beträgt jeweils etwa einen Faktor drei. Dies entspricht dem üblicherweise angegebenen Unterschied zwischen SOI-Transistoren mit einem vollständig verarmten Film und Bulk- oder Dickfilm-SOI-Transistoren [GRO90], [COL97]. Prinzipiell ist also der Ansatz aus Gleichung 6.13 anwendbar, er gilt aber immer nur für einen

[3] In Kapitel 6 wird die Schwellenspannung immer über die Tangentenmethode aus Eingangskennlinien (mit U_{DS} = ± 0,1 V) mit Hilfe der Formel $U_{th} = U_{GS} - I_{DS}/g_{m,max} - U_{DS}/2$ gewonnen.

bestimmten Temperaturbereich und $U_{th,0}$ entspricht nur dann dem *y*-Achsenabschnitt, wenn der betrachtete Fit-Bereich auch die Raumtemperatur beinhaltet.

Zwischen 400 °C und 450 °C überkreuzen sich die Schwellenspannungswerte des NMOS- und des PMOS-Transistors, so dass bei 450 °C für den NMOS eine betragsmäßig kleinere Schwellenspannung berechnet werden kann als für den PMOS (leere Symbole). Physikalisch ist eine solche Konstellation eigentlich nicht möglich, denn wenn der Transistor intrinsisch wird, d. h. wenn es keinen Unterschied mehr zwischen n- und p-dotierten Bereichen gibt, sollten die Schwellenspannungen gleich sein, sich aber nicht umkehren. Die dargestellten Werte bei 450 °C sind also nur eine Folge der Tatsache, dass die Schwellenspannung *berechnet* wurde. Da auch noch bei 450 °C bei beiden Transistoren der Drain-Source-Strom für eine betragsmäßig zunehmende Gate-Source-Spannung ansteigt, d. h. On- und Off-Zustand des Transistors weiterhin unterscheidbar sind, ist es auch noch möglich, eine Schwellenspannung zu berechnen.

Der Leckstrom I_{leck}

Der Leckstrom, gemessen als Drain-Source-Strom I_{DS} bei einer Gate-Source-Spannung von U_{GS} = 0 V ([4]) bzw. U_{GS} = ± 1 V ([5]) steigt mit der Temperatur an, was in Abbildung 6.3 verdeutlicht ist. Die Ströme sind als Beträge dargestellt, damit auch im Falle des PMOS eine logarithmische Darstellung möglich ist.

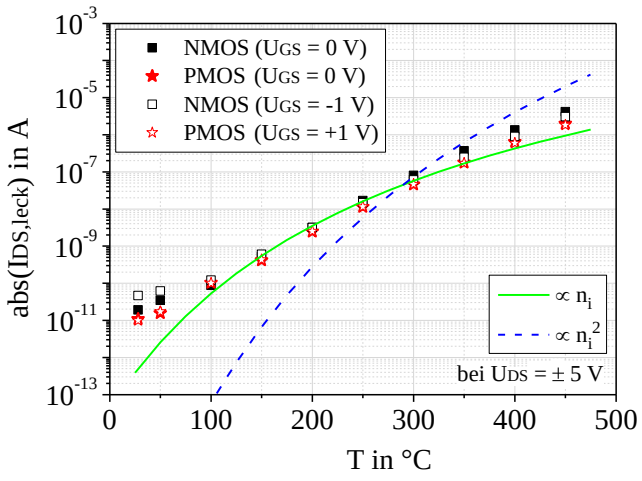


Abbildung 6.3: Betrag des Drain-Source-Stromes I_{DS} von NMOS- (Kästchen) und PMOS-Transistoren (Sterne) bei U_{GS} = 0 V(ausgefüllte Symbole) bzw. U_{GS} = ± 1 V (leere Symbole) in Abhängigkeit von der Temperatur; U_{DS} = 5 V (NMOS) und U_{DS} = -5 V (PMOS); U_S = 0 V und U_{BG} = 0 V; die Daten stammen aus den Messungen zu Abbildung 6.1

Für Temperaturen zwischen 100 °C und ungefähr 250 °C folgt der Kurvenverlauf n_i, für höhere Temperaturen tendieren die Leckströme in Richtung der Kurve, die proportional zu n_i^2 ist und auf eine Übereinstimmung mit den Messdaten bei 300 °C skaliert wurde. Dies lässt vermuten, dass für die niedrigeren Temperaturen nur der Generations-Rekombinations-Anteil

[4] zur Bestimmung des Leckstromes im Off-Zustand des Transistors
[5] zur Überprüfung, ob der Leckstrom vom pn-Leckstrom dominiert wird, oder ob der Subtreshold-Leckstrom, welcher bei U_{GS} = 0 V vorhanden sein könnte, einen wesentlichen Einfluss hat

des pn-Leckstromes eine Rolle spielt. Spätestens ab 300 °C trägt aber auch der Diffusionsanteil mit der n_i^2-Abhängigkeit zum Gesamtleckstrom bei.

Der Subtreshold-Leckstrom ist im Temperaturbereich von circa 100 °C bis 250 °C vernachlässigbar, denn die Stromwerte bei U_{GS} = ± 1, wo der Subtreshold-Leckstrom in jedem Fall vom pn-Leckstrom überdeckt werden würde, unterscheiden sich nicht von den Stromwerten bei U_{GS} = 0 V. Für höhere Temperaturen spielt er aber eine kleine Rolle, denn sowohl beim NMOS als auch beim PMOS liegen die Stromwerte bei U_{GS} = 0 V geringfügig über denen von U_{GS} = ± 1. Für Temperaturen unterhalb von 100 °C ist es genau umgekehrt. Hier sind aber die Leckströme generell höher als mit dem Verlauf von n_i zu erwarten wäre. Dies liegt daran, dass in diesem Temperaturbereich die Messung des Leckstromes sehr verrauscht ist und daher nur schwer korrekte Werte abgelesen werden können. Beim NMOS erhöht zudem ein kleiner GIDL-Effekt bei 28 °C und 50 °C den Leckstrom (vergleiche Abbildung 6.1 a)). Auch bei Colinge, der bei SOI-Transistoren im Bereich zwischen 150 °C und 300 °C eine n_i-Abhängigkeit des Leckstromes feststellen konnte, liegen die Werte bei Temperaturen unterhalb von 150 °C über der n_i-Kurve [COL97].

Betrachtet man den gesamten Verlauf des Leckstromes mit der Temperatur, wäre auch ein exponentieller Fit durch alle Messpunkte möglich. Dieses Verhalten gilt sowohl für den NMOS- als auch für den PMOS-Transistor bei Temperaturen zwischen Raumtemperatur und 450 °C. Dabei erhöht sich der Leckstrom alle 50 °C um einen Faktor von ungefähr 4,5.

Der Sättigungsstrom $I_{sätt}$

In Abbildung 6.4 ist die Temperaturabhängigkeit des Sättigungsstromes bei U_{GS} = ± 5 V und U_{DS} = ± 5 V für NMOS (a) und PMOS (b) dargestellt. Auch hier sind im Falle des PMOS die Beträge aufgetragen.

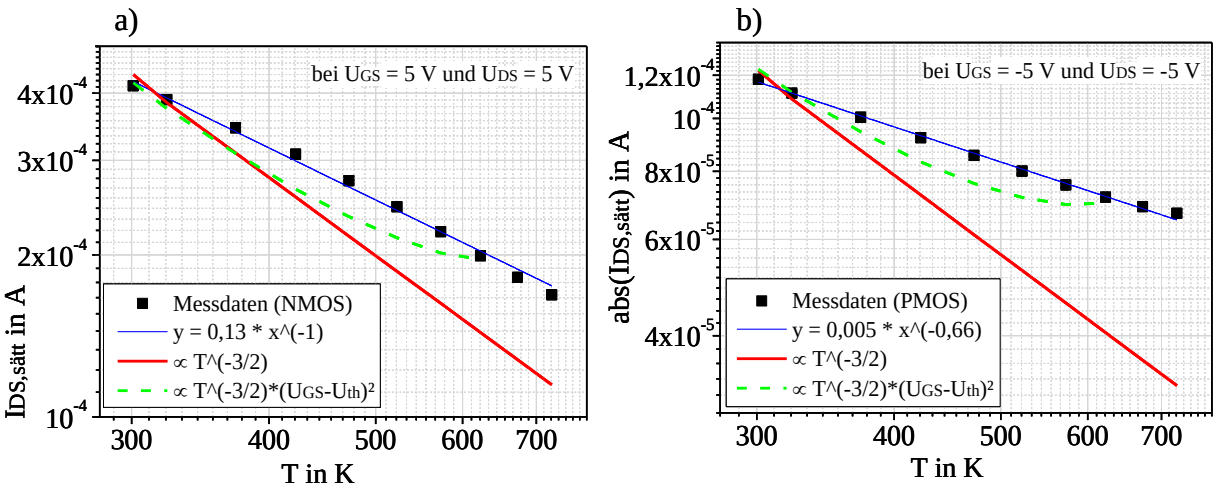


Abbildung 6.4: Drain-Source-Strom I_{DS} von NMOS- (a) und PMOS-Transistoren (b, Beträge) bei U_{GS} = 5 V (NMOS) bzw. U_{GS} = -5 V (PMOS) in Abhängigkeit von der Temperatur (hier in Kelvin); U_{DS} = 5 V und U_{DS} = -5 V (PMOS); U_S = 0 V und U_{BG} = 0 V; die Daten stammen aus den Messungen zu Abbildung 6.1

Die *x*- und die *y*-Achsen sind logarithmisch. Der Temperaturverlauf der Sättigungsströme folgt in dieser Darstellung einer Geraden und damit einem Potenzgesetz. In der doppel-

logarithmischen Darstellung mit den Temperaturen in Kelvin hat auch die Temperaturabhängigkeit der Beweglichkeit einen Geradenverlauf (dicke durchgezogene Linien). Der häufig angenommene Exponent von -3/2 für die Temperaturabhängigkeit von μ ist aber größer als die hier gefundenen Exponenten für NMOS (-1) und PMOS (-0,66). Die Kurve, in der auch die Temperaturabhängigkeit der Schwellenspannung berücksichtigt wird (gestrichelte Linien) liegt zwar näher an den Messwerten als die Kurve ohne diese Berücksichtigung (dicke durchgezogene Linien), sie folgt aber keinem Potenzgesetz, da die Schwellenspannungsänderung nicht über den gesamten Temperaturbereich gleichmäßig verläuft (siehe oben).

Der Betrag des Sättigungsstromes reduziert sich beim NMOS alle 50 °C um ungefähr 5 % und beim PMOS um etwa 10 %. Die Absolutwerte der Ströme für die beiden Transistoren unterscheiden sich um mehr als einen Faktor drei, weil die Schwellenspannung des PMOS-Transistors betragsmäßig höher ist als die des NMOS-Transistors.

Der Unterschied zwischen „On"- und „Off"-Zustand bei 450 °C, d. h. zwischen dem Leckstrom (Abbildung 6.3) und dem Sättigungsstrom (Abbildung 6.4), beläuft sich beim NMOS-Transistor auf einen Faktor von ungefähr 40, beim PMOS-Transistor auf einen Faktor von circa 35 (vergleiche dazu auch Abbildung 6.1).

Fazit

Zusammenfassend lassen sich folgende Ergebnisse aus der Charakterisierung von MOSFET-Transistoren zwischen Raumtemperatur und 450 °C festhalten:

- Der Betrag der **Schwellenspannung** nimmt mit der Temperatur ab. Der Verlauf ist sowohl für den NMOS als auch für den PMOS jeweils im Temperaturbereich von Raumtemperatur bis etwa 275 °C und von etwa 275 °C bis 400 °C annähernd linear. Die Steigung im zweiten Abschnitt ist ungefähr dreimal so groß wie im ersten Abschnitt.

- Der Betrag des **Leckstromes** steigt mit der Temperatur an und ist für NMOS und PMOS fast gleich. Bis etwa 250 °C spielt nur der Generations-Rekombinations-Anteil des pn-Leckstromes eine Rolle, für höhere Temperaturen haben auch der Diffusionsanteil des pn-Leckstromes sowie der Subtreshold-Leckstrom einen Einfluss. Der Leckstrom ändert sich alle 50 °C um einen Faktor von ungefähr 4,5 und liegt bei 450 °C in einer Größenordnung von einigen 10^{-6} A.

- Der Betrag des **Sättigungsstromes** nimmt mit der Temperatur ab. Der Kurvenverlauf folgt einem Potenzgesetz mit einem Exponenten von -1 für den NMOS und -0,66 für den PMOS. Für den NMOS-Transistor liegt alle 50 °C eine Reduzierung des Sättigungsstromes um etwa 5 % und für den PMOS-Transistor um ungefähr 10 % vor.

6.1.3 Langzeitstabilität von NMOS- und PMOS-Transistoren (ohne Versorgungsspannung)

Um festzustellen, ob sich nicht nur die Transistorkennlinien bei erhöhter Temperatur von den Ergebnissen bei Raumtemperatur unterscheiden, sondern möglicherweise dauerhafter Temperatureinfluss die Kennlinien weiter verschiebt, wurden NMOS- und PMOS-Transistoren von 1 µm und 1,6 µm Länge ohne Spannungsbelastung bei 250 °C und 350 °C gelagert[6]. Die Lagerung erfolgte in zwei Öfen, die mit Stickstoff gespült aber nicht hermetisch dicht waren. In regelmäßigen Abständen wurden Eingangskennlinien aufgenommen, um die Entwicklung der Schwellenspannung, des Leckstromes und des Sättigungsstromes mit der Zeit zu verfolgen.

Abbildung 6.5 a) zeigt die Veränderung der Schwellenspannung der NMOS-Transistoren als Mittelwert von jeweils zehn Transistoren eins Typs in Abhängigkeit von der Zeit, Abbildung 6.5 b) das entsprechende Verhalten der PMOS-Transistoren. Die Eingangskennlinien, aus denen die Berechnung der Schwellenspannung mit Hilfe der Tangentenmethode erfolgte, wurden immer bei Raumtemperatur aufgenommen.

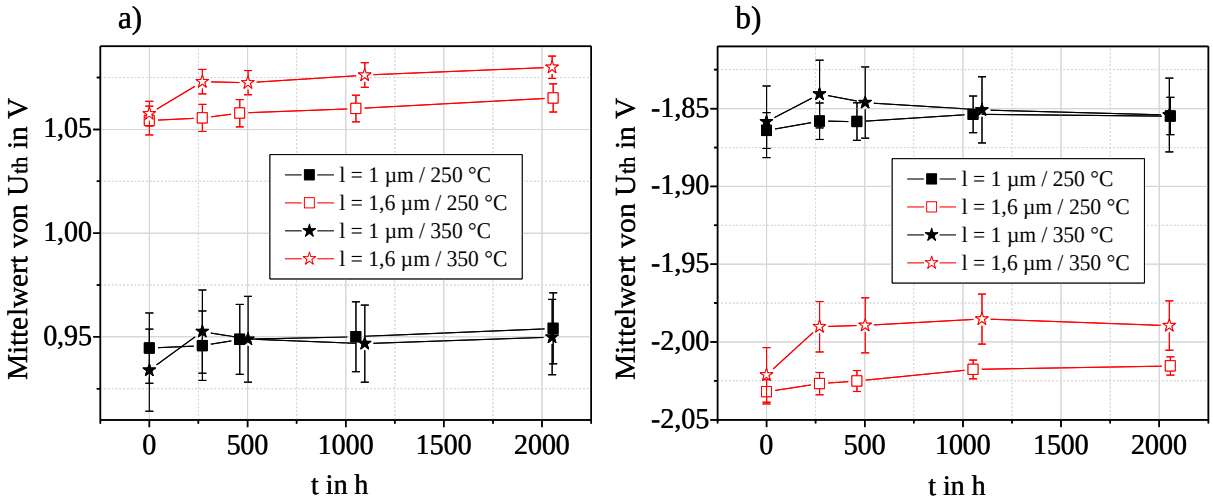


Abbildung 6.5: Schwellenspannungen von NMOS- (a) und PMOS-Transistoren (b) mit einer Kanallänge von l = 1 µm (ausgefüllte Symbole) bzw. l = 1,6 µm (leere Symbole) und einer Kanalweite w = 3,6 µm nach mehreren Stunden Lagerung bei 250 °C (Kästchen) und 350 °C (Sterne), berechnet über die Tangentenmethode aus Eingangskennlinien $I_{DS}(U_{GS})$ bei Raumtemperatur mit U_{DS} = 0,1 V (NMOS) bzw. U_{DS} = -0,1 V (PMOS); U_S = 0 V und U_{BG} = 0; Mittelwerte von jeweils zehn Transistoren eins Typs

Die Differenz der Schwellenspannungen zwischen den kürzeren und den längeren Transistoren ist auf Kurzkanaleffekte wie den Effekt der Drain-induzierten Potentialbarrierenverringerung (engl. Drain Induced Barrier Lowering oder DIBL) zurückzuführen[7]. Dieser Effekt ist für kleinere Kanallängen stärker ausgeprägt, die deshalb eine kleinere Schwellenspannung haben.

[6] Für die Langzeitstabilität von NMOS und PMOS unter Spannungseinfluss sind die Kapitel 6.1.4 (Hot Carrier) und 6.1.5 (NBTI) relevant.

[7] Beim DIBL-Effekt verringert sich durch eine hohe Drain-Source-Spannung die Potentialbarriere am Gate, wodurch der Betrag der Schwellenspannung sinkt, was vor allem bei Transistoren mit kurzen Kanallängen den Leckstrom erhöhen kann.

Bei 250 °C ändern sich die Schwellenspannungen im Bebachtungszeitraum von etwa 2000 h nur geringfügig. Im Rahmen der Messgenauigkeiten sind diese Änderungen jedoch nicht signifikant. Bei 350 °C liegt zwischen dem ersten Messpunkt bei t = 0 h und dem zweiten Messpunkt bei t = 270 h ein kleiner Anstieg der Schwellenspannung vor (< 50 mV). Der Grund dafür könnte ein Aufbrechen der bei der Abschlusstemperung am Ende der Waferprozessierung unter Wasserstoffatmosphäre abgesättigten Wasserstoffbindungen sein. Durch den erneuten Temperatureinfluss bei der hier durchgeführten Lagerung könnten einige der Bindungen wieder aufbrechen (vergleiche dazu auch Kapitel 5.2). Insgesamt sind aber keine wesentlichen Auswirkungen auf die Bauelement- und Schaltungsfunktionen zu erwarten.

Weder nimmt der Leckstrom während mehr als 2000 h bei 250 °C oder bei 350 °C zu noch nimmt der Sättigungsstrom ab. Dies gilt für NMOS- und PMOS-Transistoren beider Kanallängen. Zudem änderten sich auch die Durchbruchspannungen in den Ausgangskennlinien $I_{DS}(U_{DS})$ nicht.

Insgesamt kann also festgehalten werden, dass die vier Transistortypen mindestens 2000 h bei 250 °C und 2000 h bei 350 °C praktisch stabil sind, wenn keine Versorgungsspannung anliegt. Wie ein kombinierter Einfluss von Temperatur und Spannung auf die Transistoren wirkt, wird in den beiden nächsten Abschnitten 6.1.4 und 6.1.5 untersucht.

6.1.4 Stabilität von Transistorparametern bei hoher Drain-Source-Spannung (Hot Carrier)

Für Transistoren ist es wichtig, ein stabiles Gateoxid und eine zuverlässige Metallisierung zu haben. In der Anwendung können aber noch weitere Fehlermechanismen auftreten, die für verschiedene Schaltungszustände eines Transistors relevant sind. Dabei sind vor allem die beiden Phänomene der „heißen“ Ladungsträger (engl. „Hot Carrier“ oder HCI) und der Parameterinstabilität bei negativer Gate-Source-Spannung (engl. „Negative Bias Temperature Instability“ oder NBTI) von Bedeutung. Beide Zuverlässigkeitsaspekte werden mit Methoden untersucht, die für eine Lebensdauerbestimmung keine vollständige Zerstörung des Bauteils erfordern, sondern nur eine festgelegte Degradation der Bauteilparameter.

Hot Carrier - Effekte sind bei höheren Temperaturen weniger kritisch als bei niedrigeren Temperaturen. Ziel dieser Arbeit ist zwar die Untersuchung von Zuverlässigkeitsmechanismen bei hohen Temperaturen, da aber der H10-Prozess, auf dem diese Untersuchungen beruhen, auch für Raumtemperatur und sogar Temperaturen bis -40 °C ausgelegt ist, ist auch die Analyse von Hot Carrier - Effekten wichtig und wird deshalb hier erläutert. In erster Linie stehen in diesem Abschnitt also keine hohen Temperaturen im Vordergrund, sondern die Frage, ob und inwieweit die Anwendung der JEDEC-Messmethoden auf SOI-Transistoren möglich und sinnvoll ist.

Das Phänomen der „heißen“ Ladungsträger: physikalischer Hintergrund

„Heiße“ Ladungsträger entstehen in einem MOS-Transistor, wenn am Drain eine (betragsmäßig) sehr hohe Spannung gegenüber der Source anliegt und die Gatespannung betragsmäßig größer ist als die Schwellenspannung, damit Ladungsträger im Kanal vorhanden sind. Je höher die Drainspannung U_D ist, desto geringer ist die Ladung der Inversionsschicht am Drain. Bei Erreichen der Sättigungs-Drainspannung U_{Dsat} ist der Kanal am Drainende so eingeschnürt, dass dort keine Inversionsschicht mehr vorhanden ist (Pinch-Off, siehe Abbildung 6.6 a) für das Beispiel eines SOI-NMOS). Wenn die Drainspannung noch weiter erhöht wird, verschiebt sich der Pinch-Off-Punkt zur Source hin (Abbildung 6.6 b)). Zwischen diesem Punkt, dem Kanalende, und dem Draingebiet bildet sich ein elektrisches Feld aus. Die Ladungsträger im Kanal werden auf ihrem Weg von der Source zum Drain durch dieses elektrische Feld beschleunigt und erhalten eine zusätzliche kinetische Energie.

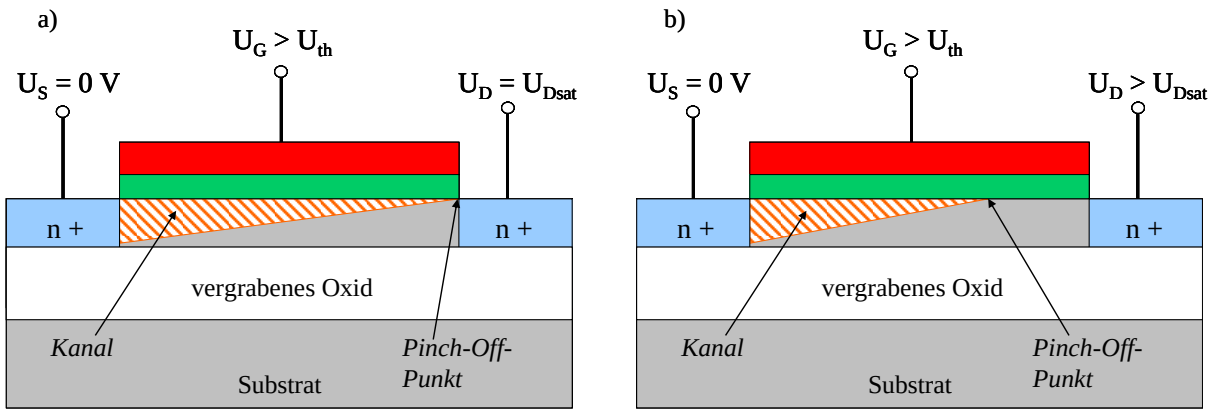


Abbildung 6.6: Schematische Darstellung zur Entstehung von „heißen“ Ladungsträgern am Beispiel eines SOI-NMOS; a) eingeschnürter Kanal an der Drain-Seite, b) Verschiebung des Pinch-Off-Punktes zur Source hin

Im thermodynamischen Gleichgewicht eines Siliziumsubstrats gilt für die Energie E eines Elektrons [LEB93]:

$$E - E_C = k \cdot T \qquad [6.18]$$

Dabei entspricht E_C der Energie der niedrigsten Energiestufe des Leitungsbands, k der Boltzmann-Konstanten und T der Substrattemperatur. Im Falle eines zusätzlich beschleunigten Elektrons ist die Energie E_e höher, so dass gilt:

$$E_e - E_C = k \cdot T_e > k \cdot T \qquad [6.19]$$

Dadurch ist auch die effektive Temperatur T_e höher als die Temperatur T und man spricht von „heißen" Elektronen. Entsprechendes gilt für „heiße" Löcher. Je höher die Drain-Source-Spannung ist, desto mehr „heiße" Ladungsträger können erzeugt werden.

Wenn diese hochenergetischen Ladungsträger genügend Energie haben, können sie die Potentialbarriere zwischen dem Siliziumsubstrat und dem Siliziumdioxid direkt überwinden [LEB93]. Für Elektronen ist die Potentialbarriere niedriger (2,8 bis 3,2 eV [CRI97], [ACO96]) als für Löcher (4,0 bis 4,7 eV [LEB93], [ACO96]). Wenn die Energie der Ladungsträger zwar hoch, aber nicht ausreichend für die direkte Überwindung dieser Barriere ist, können die durch das elektrische Feld beschleunigten Ladungsträger durch Stoßionisation Elektron-Loch-Paare erzeugen, oder durch quasi-elastischen Stoß mit einem Atom abgelenkt werden. Elektron-Loch-Paare entstehen, wenn ein beschleunigter Ladungsträger bei einem Stoß mit einem Atom ein gebundenes Elektron aus dem Valenzband ins Leitungsband hebt und im Valenzband ein Loch zurück lässt. Im Gateoxid bzw. an der Grenzfläche zwischen dem Gateoxid und dem Siliziumsubstrat können die Ladungsträger dann vorhandene Haftstellen besetzen oder neue generieren. Ladungen im Gateoxid und an der Grenzfläche zum Substrat führen zur Degradation von Transistorparametern, wie dem linearen Drain-Source-Strom[8] $I_{DS,lin}$ oder der Ladungsträgerbeweglichkeit, die sich in einer Änderung der Steilheit (Transkonduktanz) g_m der Eingangskennlinie[9] bemerkbar macht. Seltener ist auch eine Änderung der Schwellenspannung U_{th} zu beobachten.

Die Energiegewinnung von Elektronen, die in einem elektrischen Feld stark beschleunigt werden, wird durch das „Lucky Electron Model" beschrieben [HU85], [JED122]. Dieses Modell ist aber für eine vollständige Darstellung der Hot Carrier - Phänomene nicht ausreichend. Die Bedingungen, bei denen die größte Hot Carrier - Schädigung auftritt, sind nämlich von der Transistor-Architektur und vor allem von der Kanallänge abhängig.

Ob Elektronen oder Löcher in das Gateoxid injiziert werden, Haftstellen besetzen oder generieren, hängt von der Potentialdifferenz zwischen Gate und Drain und vom Transistortyp ab. Ist beim **NMOS-Transistor** das Potential am Gate viel kleiner als das Potential am Drain, werden vor allem „heiße" Löcher in Richtung des Gateoxids abgelenkt [ACO96]. Die Größe des Löcherstromes zum Substrat korreliert mit der Zahl an Löchern, die zum Gateoxid abgelenkt werden (als Folge des Avalanche-Effektes) und ist damit ein Maß für die Veränderung des Transistors durch „heiße" Ladungsträger [JED28A]. Mit kleiner werdendem

[8] $I_{DS,lin}$ ist der Strom im linearen Bereich der Transistor-Eingangskennlinie $I_{DS}(U_{GS})$. Für eine Schwellenspannung von U_{th} = 1 V kann so zum Beispiel $I_{DS,lin}$ bei U_{GS} = 2 V betrachtet werden.

[9] Die Steilheit bzw. Transkonduktanz g_m ist der Quotient aus δI_D und δU_{GS} bei U_{DS} = konstant im linearen Bereich der Eingangskennlinie.

Unterschied zwischen Drain- und Gate-Potential nimmt die Injektion „heißer“ Elektronen in das Gateoxid zu. Aufgrund ihrer größeren Beweglichkeit, der größeren mittleren freien Weglänge, der größeren Stoßionisationsrate und der kleineren Barrierenhöhe ist es für Elektronen einfacher als für Löcher, in das Gateoxid zu gelangen. Bei einer bestimmten Kombination von Gate- und Drain-Potential wird dann der Substratstrom maximal. Dabei ist die Stoßionisationsrate sehr hoch und sowohl „heiße“ Löcher als auch „heiße“ Elektronen gelangen in das Gateoxid. Diese Kombination von Gate- und Drain-Potential liegt in etwa bei $U_G \approx U_D/2$ [CRI97]. Wegen der Elektroneninjektion ins Gateoxid, wo die Elektronen Haftstellen besetzen oder neue generieren können, steigt die Schwellenspannung an und der lineare Drain-Source-Strom sinkt. Wird das Potential am Gate gleich groß oder sogar größer als am Drain, überwiegt die Injektion von Elektronen ins Gateoxid [CRI97].

Bei Technologien mit einer Kanallänge von $l < 0{,}35$ µm ist es vor allem der Fall $U_G \approx U_D$, bei dem die maximale Veränderung der Transistorparameter auftritt. Für Technologien mit einer Kanallänge von $l > 0{,}35$ µm gilt meistens $U_G \approx U_D/2$ als kritischste Kombination. Das genaue Verhältnis von U_G und U_D kann bestimmt werden, indem bei einer bestimmten Drainspannung die Gatespannung ermittelt wird, bei der der maximale Substratstrom auftritt.

Für den **PMOS-Transistor** ist die Injektion von Löchern und Elektronen genau umgekehrt zum NMOS-Transistor. Für eine betragsmäßig viel kleinere Gatespannung im Vergleich zur Drainspannung überwiegt die Injektion „heißer“ Elektronen ins Gateoxid [ACO96]. Auch hier gibt es eine Kombination von Drain- und Gatespannung, bei der die maximale Parameteränderung auftritt. Um diese zu bestimmen, wird bei einer festen Drainspannung die Gatespannung ermittelt, bei der der Gatestrom maximal wird [JED60]. Wegen der durch die abgelenkten Elektronen größeren Anzahl an negativ geladenen Defekten oder Haftstellen im Gateoxid sinkt der Betrag der Schwellenspannung des PMOS-Transistors. Der lineare Drainstrom steigt an, d.h. der Transistor wird „stärker“. Aus diesem Grund wurde lange Zeit die Parameterdegradation von PMOS-Transistoren durch den Einfluss „heißer“ Ladungsträger als unkritisch eingestuft und vor allem die Veränderungen an NMOS-Transistoren untersucht.

Für Technologien mit kleinerer Kanallänge spielen aber auch positive Grenzflächen-Oxidladungen und die Injektion von Löchern eine Rolle [IOA06]. Wie auch bei den NMOS-Transistoren kann dann in dem Fall, in dem Drain- und Gatepotential ungefähr gleich groß sind, die größte Parameteränderung beobachtet werden.

Für jeden Transistor gibt es also einen Arbeitspunkt, bei dem die Kombination aus dem lateralen elektrischen Feld und dem vertikalen elektrischen Feld zu maximaler Transistorschädigung führt. Es gibt Möglichkeiten, die Entstehung von „heißen“ Ladungsträgern bzw. ihre Energiegewinnung zu beeinflussen. Beispielsweise kann mit einer so genannten „Lightly Doped Drain (LDD)“ - Zone am Drain das elektrische Feld und damit die Ladungsträgerbeschleunigung reduziert werden [OGU80], [ACO96].

Hot Carrier - Stress ist zudem lokaler Degradationsmechanismus, denn die Schädigung tritt hauptsächlich im Drainbereich auf. Sie ist nicht reversibel, das heißt auch nach Abschalten der Stressspannungen bleiben die Transistorparameter verändert.

Die Degradation der Transistorparameter durch „heiße“ Ladungsträger nimmt mit steigender Temperatur ab. Für höhere Temperaturen ist die mittlere freie Weglänge, d.h. die Strecke, die Ladungsträger ohne einen Stoß zu erfahren, zurücklegen können, kleiner. Die Ladungsträger können weniger beschleunigt werden als bei niedrigeren Temperaturen, erhalten somit auch weniger Energie und erzeugen dadurch weniger Elektron-Loch-Paare.

Die obigen allgemeinen Erklärungen der Hot Carrier - Phänomene beziehen sich auf den Einfluss „heißer“ Ladungsträger auf das Gateoxid bzw. die Grenzfläche zwischen dem Siliziumsubstrat und dem Gateoxid. Bei SOI-Transistoren können entsprechende Phänomene aber auch an der Grenzfläche zum vergrabenen Oxid auftreten. Dass die Wahl des Back-Gate-Potentials einen maßgeblichen Einfluss auf die Veränderung der Transistorparameter hat, zeigt sich bei den Untersuchungen an den H10-Transistoren (siehe Abschnitt „Hot Carrier - Effekte beim H10-Prozess“).

Die Degradation durch „heiße“ Ladungsträger ist nicht nur ein Phänomen, das einen Transistor als einzelnes Bauelement betreffen kann, sondern auch in einer Schaltung möglich ist, wenn an einem Transistor eine hohe Drain-Source-Spannung anliegt. In analogen Schaltungen zum Beispiel ist es vor allem die Geschwindigkeit der Schaltung, die durch „heiße“ Ladungsträger verringert wird.

Messmethode und Berechnung der Lebensdauer eines Transistors mit Hot Carrier - Schädigung

Die Degradation der Transistorparameter durch Hot Carrier - Effekte, im Englischen auch „Hot Carrier Integrity (HCI)“, „Hot Carrier Stress (HCS)“ oder „Channel Hot Carrier (CHC)“ genannt, kann mit Hilfe der JEDEC-Standards JESD28-A, JESD28-1 und JESD60A [JED28A], [JED28], [JED60] untersucht werden. Das Vorgehen bei der Bestimmung der Messbedingungen, der Messung selbst und der Auswertung der Ergebnisse wird im Folgenden beschrieben.

1) Wie im ersten Teil des Abschnitts 6.1.4 erläutert, treten Hot Carrier - Degradationen bei großen Drain-Source-Spannungen U_{DS} auf. Die maximal erlaubte Drain-Source-Stressspannung $U_{DS,stress}$, bei der laut JEDEC-Standards Hot Carrier - Zuverlässigkeitstests durchgeführt werden können, ist beim NMOS die Durchbruchspannung und beim PMOS 80 % der Durchbruchspannung. Die Durchbruchspannung erhält man aus dem Ausgangskennlinienfeld $I_{DS}(U_{DS})$. Transistoren, die für die Bestimmung dieses Kennlinienfeldes verwendet wurden, dürfen später nicht für die Stressmessungen benutzt werden, da sie durch diese erste Messung vorbelastet sind. Für eine genaue Bestimmung der Hot Carrier - Degradation sind Messungen mit mindestens drei verschiedenen Drain-Source-Stressspannungen durchzuführen.

2) Die zur jeweiligen Drain-Source-Stressspannung $U_{DS,stress}$ passende Gate-Source-Stressspannung $U_{GS,stress}$ wird, wie bereits erläutert, durch den maximalen Substratstrom $I_{B,max}$ (NMOS) bzw. den maximalen Gatestrom $I_{G,max}$ (PMOS) ermittelt. Da der Substratstrom bei den hier vorliegenden SOI-Transistoren nicht bestimmt werden kann, weil der Siliziumfilm nicht separat angeschlossen ist, müsste man für die Ermittlung der kritischsten $U_{DS,stress}$-$U_{GS,stress}$-Kombination Hot Carrier - Messungen bei verschiedenen Spannungskombinationen durchführen und den Fall der schlimmsten Degradation heraussuchen, was aber sehr aufwendig ist. Einfacher ist es, den Filmstrom eines HGate-Transistors, dessen Kanallänge und Kanalweite möglichst denen des eigentlich zu messenden Transistors entsprechen, als Alternative zum Substratstrom zu messen [WAN93].

3) Hat man für mindestens drei Drain-Source-Stressspannungen die entsprechenden Gate-Source-Stressspannungen gefunden, die jeweils zu maximaler Hot Carrier - Degradation führen, werden diese Spannungen für längere Zeit an den Transistor angelegt und die Änderung der Transistorparameter beobachtet. Dazu wird vor dem Anlegen des Stresses und

dann in diskreten Zeitabständen eine Eingangskennlinie aufgenommen. Beim NMOS folgt eine typische Hot Carrier - Degradation einem Potenzgesetz (vergleiche Abbildung 6.7 a)). Beim PMOS kommt für die zeitliche Degradation entweder auch ein Potenzgesetz oder ein logarithmischer Zusammenhang (vergleiche Abbildung 6.7 b)) in Frage. $|Y(t)|$ ist dabei die prozentuale Änderung des betrachteten Parameters $P(t)$ in Bezug auf diesen Parameter $P(0)$ zum Zeitpunkt $t = 0$ s. Es werden immer Beträge betrachtet, um auch die negativen Änderungen logarithmisch darstellen zu können.

$$Y(t) = \frac{P(t) - P(0)}{P(0)} \cdot 100 \qquad [6.20]$$

In allen Fällen empfiehlt es sich, logarithmisch äquidistante Zeitabstände zwischen den Messungen zu wählen.

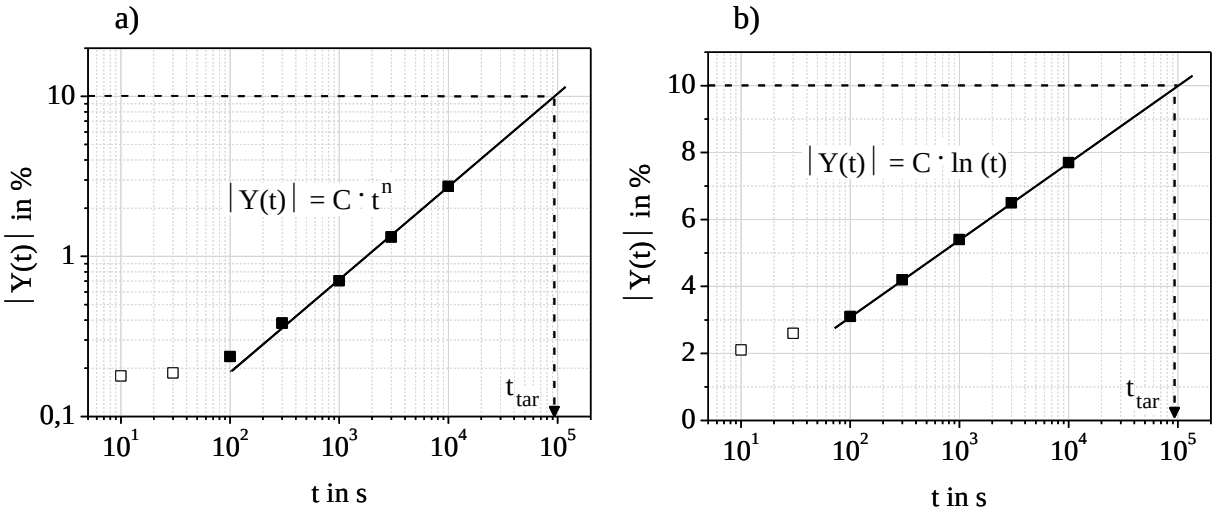


Abbildung 6.7: Beispiel für die prozentuale Änderung eines Transistor-Parameters $|Y(t)|$ in Abhängigkeit von der Zeit als Potenzgesetz (a) oder als logarithmischer Zusammenhang (b); die Änderung wird auf den Messpunkt bei $t = 0$ s bezogen; mit einer Extrapolation lässt sich die Ausfallzeit t_{tar} ermitteln, die in der Messzeit von 10.000 s nicht erreicht wurde

4) Die Degradation der Transistorparameter wird über der Zeit beobachtet. Wenn eine der Messgrößen eine bestimmte relative oder absolute Änderung erfahren hat, wird die Messung abgebrochen und der Transistor als defekt angesehen. Der Zeitpunkt, bei dem die festgelegte Degradation überschritten wird, ist die Ausfallzeit t_{tar} ([10]). Die JEDEC-Standards JESD28-A und JESD60A empfehlen, für den linearen Drain-Source-Strom $I_{DS,lin}$ und für die maximale Steilheit g_m der Eingangskennlinie eine Änderung von jeweils höchstens 10 % zu akzeptieren. Für die maximal erlaubte Änderung der Schwellenspannung werden 50 mV empfohlen. Der Parameter, der als erster die festgelegte Akzeptanzgrenze überschreitet, bestimmt die Ausfallzeit. Wenn das Ausfallkriterium nicht erreicht wird, aber genügend Messpunkte vorliegen, kann die Ausfallzeit auch über eine Extrapolation ermittelt werden (vergleiche auch Abbildung 6.7). Die Standards empfehlen, die Extrapolation erst ab einer Messzeit von 100 s durchzuführen, da sonst Auflösungsprobleme bei der Messung zu Ungenauigkeiten in der Bestimmung der Parameteränderungen führen können.

[10] Der Index „tar“ steht für das englische Wort „target“, was Ziel bedeutet. In den JEDEC-Standards wird die Ausfallzeit bei HCI- und NBTI-Messungen mit t_{tar} bezeichnet, weshalb auch hier diese Variante gewählt wird.

5) Wenn für verschiedene $U_{GS,stress}$ / $U_{DS,stress}$ - Kombinationen jeweils mehrere Exemplare einer Transistorvariante vermessen wurden, lässt sich aus den gewonnen Ausfallzeiten die Lebensdauer des Transistortyps berechnen. Der JEDEC-Standard JESD28-1 schlägt dazu drei Möglichkeiten vor. Bei zwei dieser Methoden geht der Substratstrom in die Berechnung mit ein, weshalb für die Lebensdauerbestimmung von SOI-Transistoren nur die dritte Variante geeignet ist. Bei dieser so genannten „Drain-source voltage acceleration method" wird von einem exponentiellen Zusammenhang zwischen der Ausfallzeit t_{tar} und dem Kehrwert der Drain-Source-Stressspannung $U_{DS,stress}$ ausgegangen, der in Gleichung 6.21 dargestellt ist.

$$t_{tar} = t_0 \cdot \exp\left(\frac{B}{U_{DS,stress}}\right) \qquad [6.21]$$

Dabei sind t_0 und B Fit-Parameter. Eine Auftragung von $\ln(t_{tar})$ über $1/U_{DS,stress}$ ergibt eine Gerade mit der Steigung B und dem y-Achsenabschnitt $\ln(t_0)$. Aus der Geradengleichung lässt sich dann die Lebensdauer für die maximale Betriebsspannung $U_{DS,max}$ des Transistors ermitteln.

Um mögliche Zuverlässigkeitsprobleme durch „heiße" Ladungsträger zu detektieren, sollten Untersuchungen an allen Transistorvarianten einer Technologie durchgeführt werden. Die stärkste Degradation tritt aber bei minimaler Kanallänge auf, da hier das laterale elektrische Feld zwischen Source und Drain am größten ist. Die Kanalweite hat weniger Einfluss auf die Degradation durch Hot Carrier. Zur Vermeidung von Schmalkanaleffekten empfehlen die JEDEC-Standards, Transistoren mit einer Kanalweite von mehr als 10 µm zu verwenden [JED28A].

Hot Carrier - Effekte beim H10-Prozess

Die Untersuchungen der Hot Carrier - Effekte des H10-Prozesses wurden an NMOS- und PMOS-Transistoren der Kanallängen l = 1 µm und l = 1,6 µm durchgeführt. Die Kanallänge von 1 µm ist die kleinste erlaubte Kanallänge der digitalen Transistoren ($U_{DS,max}$ = 3,3 V) und gemäß der Theorie am anfälligsten für Hot Carrier - Schädigung. Die Kanallänge von 1,6 µm entspricht der Kanallänge der analogen Minimaltransistoren ($U_{DS,max}$ = 5 V). Bezüglich der Kanalweite wurden Transistoren von w = 3,6 µm und w = 36 µm getestet (siehe auch [VOL11]), wobei keine Unterschiede in der Veränderung der Transistorparameter festgestellt werden konnten.

In Abbildung 6.8 sind Lebensdauerplots für die beiden NMOS-Transistortypen (l = 1 µm und l = 3,6 µm) und die beiden PMOS-Transistortypen (l = 1 µm und l = 3,6 µm) dargestellt. Die Weite war jeweils w = 36 µm. Die Messungen erfolgten bei Raumtemperatur (RT). Weitere Experimente mit einzelnen Transistoren bei einer Temperatur von -40 °C zeigten wie erwartet eine geringere Lebensdauer als bei Raumtemperatur. Für eine genaue Analyse der Degradationen bei tieferen Temperaturen wäre aber eine größere Statistik nötig, die im Rahmen dieser Arbeit nicht realisiert werden konnte. Hier ging es nur prinzipiell darum, wie sich Hot Carrier - Degradationen auf die H10-Transistoren auswirken und auf welche Aspekte geachtet werden muss. Deshalb wurde auch auf Messungen bei höheren Temperaturen verzichtet, bei denen nur geringe Auswirkungen der hohen Drain-Source-Spannung zu erwarten wären.

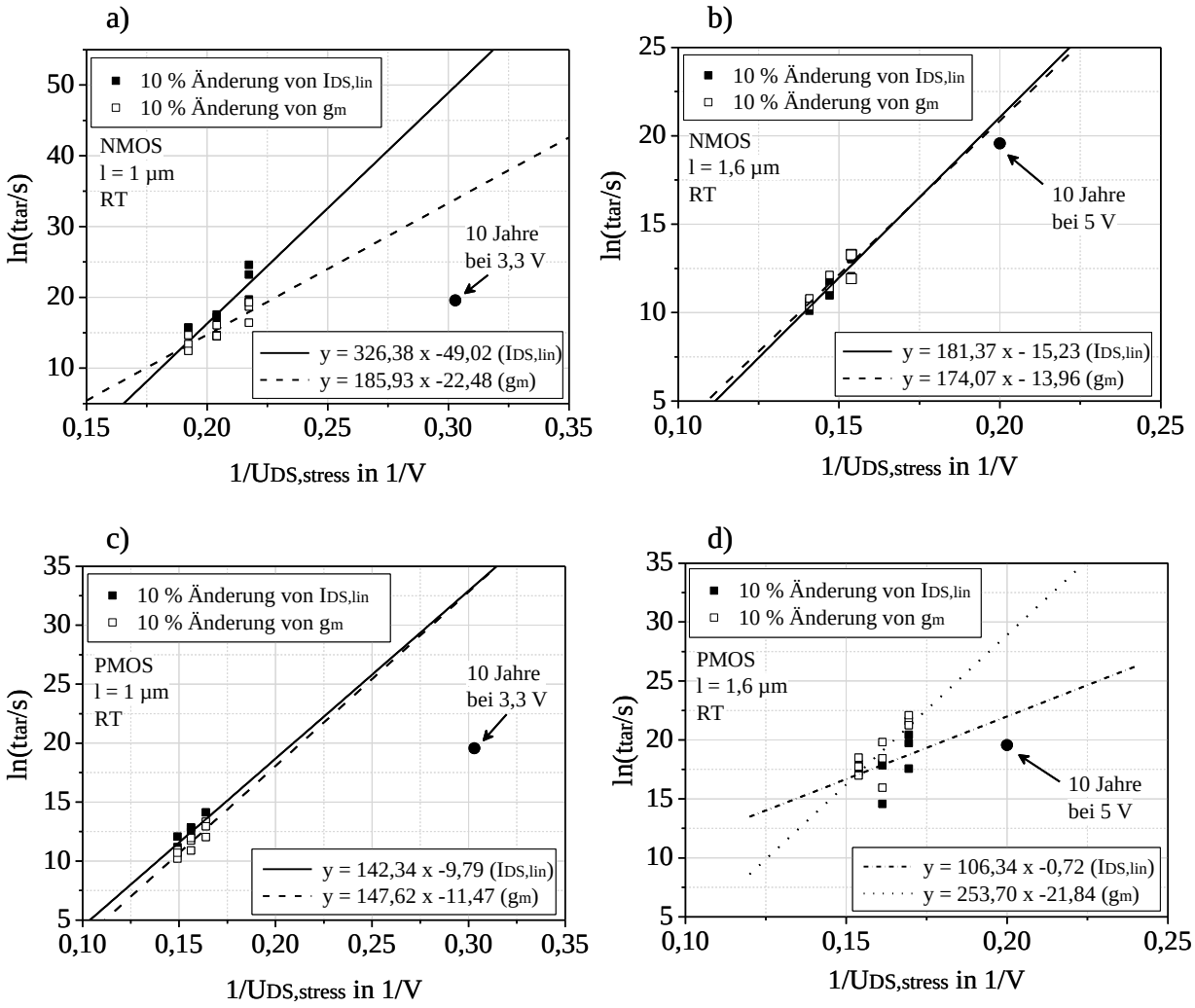


Abbildung 6.8: Lebensdauerplots nach Gleichung 6.8 für die NMOS-Transistoren (a) l = 1 µm und (b) l = 1,6 µm und die PMOS-Transistoren (c) l = 1 µm und (d) l = 1,6 µm; alle haben eine Weite von w = 36 µm; für jeden Transistortyp wurden je drei Messungen bei drei verschiedenen $U_{DS,stress}$ - $U_{GS,stress}$ - Kombinationen bei Raumtemperatur durchgeführt; Potentiale an Source und Back-Gate $U_S = U_{BG} = 0$ V; die schwarzen Kästchen stellen die 10 %-ige Änderung von $I_{DS,lin}$, die leeren Kästchen die 10%-ige Änderung von g_m dar

Für alle Transistoren, die zu den Ergebnissen in Abbildung 6.8 beigetragen haben, wurden jeweils drei Messungen bei drei Stressspannungen durchgeführt. Bei den NMOS-Transistoren lag dabei das Potential an der Source auf 0 V und war am Drain positiv ($U_{DS} > 0$ V) und bei den PMOS-Transistoren wurde mit 0 V am Drain und einem positiven Potential an der Source ($U_{DS} < 0$ V) gearbeitet. In beiden Fällen lag das Back-Gate auf 0 V. Die Ausfallzeiten t_{tar} wurden wie in Abbildung 6.7 a) demonstriert ermittelt und die Lebensdauer jeweils mit Gleichung 6.21 berechnet.

Die Ausfallzeiten beziehen sich auf eine 10 %-ige Änderung von $I_{DS,lin}$ bzw. g_m, wobei für die jeweilige Bestimmung oder Extrapolation der Ausfallzeit sowohl bei den NMOS- als auch bei den PMOS-Transistoren die Anwendung eines Potenzgesetzes wie in Abbildung 6.7 a) sinnvoll war. Es ist zu beachten, dass beim NMOS-Transistor die Änderungen negativ, beim PMOS-Transistor positiv waren. Die Änderung der Schwellenspannung war in allen Fällen sehr gering und wurde deshalb für eine Lebensdauerextrapolation nicht herangezogen.

Der Faktor zwischen $U_{DS,stress}$ und $U_{GS,stress}$ wurde für die NMOS-Transistoren über den Filmstrom von HGate-Transistoren der Länge l = 1 µm ermittelt. Er betrug in diesem Fall $U_{GS,stress}$ / $U_{DS,stress}$ = 0,5. Da keine HGate-Transistoren der Länge l = 1,6 µm zur Verfügung standen, wurde dieser Faktor sowohl für die NMOS-Transistoren mit Kanallänge l = 1 µm als auch für die NMOS-Transistoren mit l = 1,6 µm verwendet. Exemplarische Messungen mit anderen $U_{DS,stress}$ - $U_{GS,stress}$ - Kombinationen für die NMOS-Transistoren mit l = 1,6 µm zeigten aber, dass auch hier 0,5 der optimale Faktor von $U_{GS,stress}$ / $U_{DS,stress}$ war. Es liegt also der Fall $U_G = U_D/2$ vor (siehe oben).

Für die PMOS-Transistoren konnte das Verhältnis $|U_{GS,stress}| / |U_{DS,stress}|$ zu 0,1 beim Transistor der Länge l = 1µm und zu 0,3 beim Transistor der Länge l = 1,6 µm bestimmt werden. Auch hier stimmt das Verhältnis gut mit den Angaben in der Literatur für Technologien mit minimalen Kanallängen von l > 0,35 µm überein (siehe oben).

Die in Abbildung 6.8 gezeigten Ergebnisse sind sehr unterschiedlich. In allen Fällen ist die Lebensdauer bei der maximalen Einsatzspannung ($U_{DS,max}$ = 3,3 V bei den 1 µm langen Transistoren und $U_{DS,max}$ = 5 V bei den 1,6 µm langen Transistoren) größer als 10 Jahre. Für den längeren NMOS-Transistor und den kürzeren PMOS-Transistor liegen die gemessenen Werte eng beieinander. Zudem unterscheiden sich die Zeiten, in denen eine 10%-ige Änderung von $I_{DS,lin}$ erreicht wurde, nicht wesentlich von denen, in denen eine 10%-ige Änderung von g_m vorlag. Die Änderungen für $I_{DS,lin}$ und g_m folgen dabei einem Potenzgesetz. Für den kürzeren NMOS-Transistor ist die Streuung größer und g_m scheint schneller zu degradieren als $I_{DS,lin}$. Beim längeren PMOS-Transistor schwanken die einzelnen Messwerte sehr stark. Der Grund dafür ist weniger physikalischer Natur als auf die Tatsache zurückzuführen, dass insgesamt nur eine geringe Statistik vorliegt. Eine größere Anzahl an Messungen war aber im Rahmen dieser Arbeit nicht möglich.

Einfluss des Back-Gates

Wie erläutert wurden die PMOS-Transistoren für Abbildung 6.8 statt mit einem hohen Potential am Drain mit einem hohen Potential an der Source und dafür 0 V am Drain gestresst, so wie sie auch in einer Schaltung betrieben werden würden. Dabei lag das Back-Gate auf 0 V. In einem zusätzlichen Experiment wurde nun der Einfluss des Back-Gates untersucht. Abbildung 6.9 zeigt das Ergebnis als Änderung des linearen Drain-Source-Stromes in Abhängigkeit von der Zeit.

Für die erste Messung (Kästchen) wurden die Potentiale symmetrisch zum NMOS gewählt, d. h. negative Potentiale an Drain und Gate, 0 V an der Source und 0 V am Back-Gate. Bei dieser Kurve fällt auf, dass sich der erste Messpunkt nach Beginn des Stresses (bei t = 10 s) schon um mehr als 20 % vom Messpunkt bei t = 0 s unterscheidet[11]. Für die zweite Messung (Kreise) wurde dann der normale Schaltungsfall gewählt (wie in Abbildung 6.8), d. h. 0 V am Drain und ein positives Source-Potential. Hierbei fällt auf, dass die Messwerte deutlich unter denen aus der ersten Messung liegen. Der Unterschied der beiden Messungen liegt in den Potentialdifferenzen zum Back-Gate. Während in der ersten Messung U_{DBG} = -5,5 V ist, gilt für die zweite Messung U_{DBG} = 0 V. Deshalb wurde für eine dritte Messung (Sterne) im Vergleich zur zweiten Messung das Potential am Back-Gate erhöht, so dass nun die gleichen Potentialdifferenzen wie bei der ersten Messung vorliegen. Als Konsequenz befinden sich nun

[11] Es ist zu beachten, dass vorher eine Änderung von 10 % als Ausfallkriterium festgelegt wurde, so dass dieser Transistor nun schon nach weniger als 10 s als defekt angesehen werden müsste.

die gemessenen Änderungen des linearen Drain-Source-Stromes in Höhe der Werte aus der ersten Messung[12].

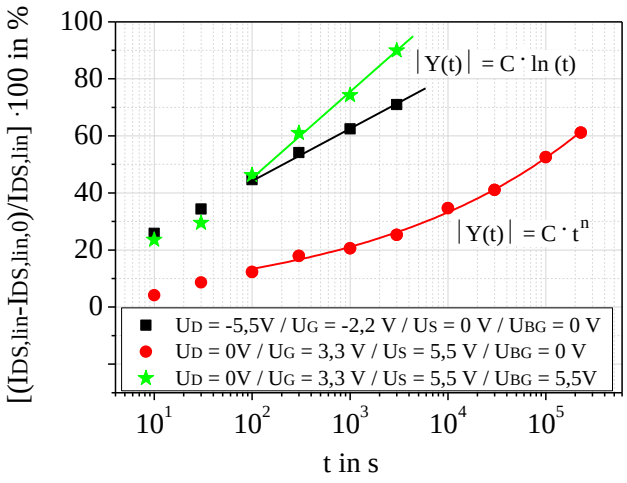


Abbildung 6.9: Prozentuale Änderung von $I_{DS,lin}$ bezogen auf den Messwert $I_{DS,lin,0}$ nach t = 0 s in Abhängigkeit von der Zeit für drei verschiedene Kombinationen der Potentiale an Drain, Gate, Source und Back-Gate; Messungen an PMOS-Transistoren der Länge l = 1 µm und der Weite w = 36 µm bei Raumtemperatur[13]

Weiterhin fällt auf, dass die prozentuale Änderung von $I_{DS,lin}$ für die beiden Fälle, in denen der Messpunkt nach 10 s schon mehr als 20 % über dem Wert von $I_{DS,lin}$ bei t = 0 s liegt, logarithmisch mit der Zeit verläuft. Die Degradation im eigentlichen Schaltungsfall (Kreise) folgt aber eher einem Potenzgesetz. Ein Potenzgesetz liegt immer dann vor, wenn nur ein Mechanismus dominant ist, in diesem Fall ist es die Schädigung durch Defekte an der Grenzfläche vom Siliziumfilm zum Siliziumdioxid. Wenn die Drain-Back-Gate-Spannung U_{DBG} nun aber zunimmt, wie es für die erste und die dritte Messung (Kästchen und Sterne) der Fall ist, können zusätzlich Defekte an der Grenzfläche zwischen dem Siliziumfilm und dem vergrabenem Oxid entstehen und den Verlauf der Degradation ändern [CRI97]. Für diesen Fall, in dem Hot Carrier - Schädigung an zwei Grenzflächen (Si-SiO_2 und Si-BOX) auftritt, liegt kein Potenzgesetz sondern ein logarithmischer Zusammenhang vor (vergleiche dazu auch Abbildung 6.7 und den JEDEC-Standard JESD60A).

Zusätzlich ist anzumerken, dass im betrachteten Zeitfenster zwar die Änderung von $I_{DS,lin}$ in der ersten und der dritten Messung wesentlich stärker ist als in der zweiten Messung, in allen drei Fällen aber die Degradation des Parameters g_m ähnlich ist. Dies bedeutet, dass die Änderung der Ladungsträgerbeweglichkeit in allen drei Fällen vergleichbar ist und somit auch die Schädigung, die an der oberen Grenzfläche (Si-SiO_2) auftritt. Für die erste und die dritte Messung kommt also die Schädigung an der Grenze vom Siliziumfilm zum vergrabenem Oxid (Si-BOX) *zusätzlich* hinzu, d. h. insgesamt ist die Schädigung durch „heiße“ Ladungsträger in der ersten und dritten Messung größer als in der zweiten Messung.

[12] Dass die Messpunkte der ersten und der dritten Messung (Kästchen und Sterne) nicht genau übereinander liegen, ist statistisch bedingt, da in jedem Fall ein anderes Exemplar der 1 µm - PMOS - Transistoren gestresst wurde.

[13] Die Potentiale an Drain und Gate sowie das Verhältnis $U_{GS,stress}$ / $U_{DS,stress}$ unterscheiden sich von denen aus Abbildung 6.8, da die Tests für Abbildung 6.9 an Transistoren auf einem anderen Wafer aus einer anderen Charge durchgeführt wurden.

Eine vergleichbare Untersuchung für den NMOS-Transistor wurde nicht vorgenommen, da in diesem Fall ein negatives Potential an das Back-Gate angelegt werden müsste, was aber schaltungstechnisch nicht vorgesehen ist.

Hot Carrier - Effekte beim H10-Prozess im Vergleich zu anderen Technologien

Die Richtung, in die die Parameteränderungen verlaufen, hängt, wie bereits angedeutet, von der Technologie ab. Tabelle 6.1 gibt einen Überblick über die betragsmäßige Änderung des linearen Drain-Source-Stromes $I_{DS,lin}$, der Schwellenspannung U_{th} und der Steilheit der Eingangskennlinie g_m, wie sie in der Literatur zu finden ist [LEB93], [IOA06] und wie sie durch Messungen für den H10-Prozess bestimmt werden konnte. Die Transistoren des H10-Prozesses verhalten sich demnach so, wie es für entsprechende Technologien mit einer Strukturgröße von 1 µm zu erwarten wäre.

	NMOS			**PMOS**		
Technologie	$l > 0{,}35$ µm	$l < 0{,}35$ µm	H10	$l > 0{,}35$ µm	$l < 0{,}35$ µm	H10
$\|U_{th}\|$	steigt	steigt	steigt	sinkt	steigt	sinkt
$\|I_{DS,lin}\|$	sinkt	sinkt	sinkt	steigt	sinkt	steigt
$\|g_m\|$	sinkt	sinkt	sinkt	steigt	sinkt	steigt

Tabelle 6.1: Überblick über die Änderung der Transistorparameter linearer Drain-Source-Strom $I_{DS,lin}$, Schwellenspannung U_{th} und Steilheit der Eingangskennlinie g_m durch Hot Carrier - Einfluss für NMOS- und PMOS-Transistoren verschiedener Technologien

Fazit

Aus den Messungen zum Einfluss von „heißen“ Ladungsträgern von 3,3 V - und 5,5 V - Transistoren kann man entnehmen, dass eine betragsmäßig hohe Drain-Source-Spannung einen Einfluss auf den linearen Drain-Source-Strom $I_{DS,lin}$ und die Steilheit der Eingangskennlinie g_m hat. Für den PMOS-Transistor ist es dabei wichtig, den Einfluss des Back-Gates zu beachten. Bei einer großen Potentialdifferenz zwischen Drain und Back-Gate erfolgt eine Hot Carrier - Schädigung nicht nur an der Grenzfläche zwischen Film und Gateoxid (Si-SiO_2), sondern auch an der Grenzfläche zwischen Film und vergrabenem Oxid (Si-BOX). Dieser Aspekt ist für SOI-Transistoren sehr wichtig, zumal er in den für Bulk-Transistoren ausgerichteten JEDEC-Standards nicht explizit erwähnt wird. Insgesamt sind die Änderungen beim PMOS-Transistor aber nicht kritisch, weil der Transistor „stärker“ wird, wie es für Technologien mit 1 µm minimaler Kanallänge zu erwarten ist. Deshalb kann man insgesamt folgern, dass bei Raumtemperatur Hot Carrier - Schädigungen keinen Einfluss auf eine zuverlässige Funktionalität der beiden H10-Transistortypen haben und somit auch bei 250 °C keine Auswirkungen zu erwarten sind.

6.1.5 Stabilität von Transistorparametern bei hoher Gate-Source-Spannung (Negative Bias Temperature Instability)

Neben der Degradation eines Transistors durch die Injektion „heißer“ Ladungsträger gibt es einen zweiten Effekt, der Transistorparameter dauerhaft schädigen kann. Es handelt sich dabei um die Parameterinstabilität bei negativer Gate-Source-Spannung (engl. „Negative Bias Temperature Instability“ oder NBTI). NBTI tritt im Gegensatz zu Hot Carrier - Effekten nicht bei hoher Drain-Source-Spannung, sondern bei hoher Gate-Source-Spannung auf und ist zudem kritischer je höher die Temperatur wird. Es ist also nicht das laterale elektrische Feld E_{lat} zwischen Source und Drain, sondern das vertikale elektrische Feld E_{ver} zwischen Gate und Kanal, das den Grad der Schädigung beeinflusst. Der NBTI-Effekt betrifft vor allem PMOS-Transistoren.

Physikalische Ursachen für die Transistorschädigung durch eine negative Gate-Source-Spannung

Das Phänomen der Parameterinstabilität bei negativer Gate-Source-Spannung ist seit den 1960er Jahren bekannt und seitdem eines der größten und umstrittensten Zuverlässigkeitsprobleme der CMOS-Technologie [GRA11]. Dabei sind die physikalischen Ursachen noch nicht vollständig geklärt, und deshalb gibt es auch bei den Messmethoden und der Auswertung der Messdaten unterschiedliche Ansätze. Im Folgenden soll ein kurzer Überblick über mögliche Erklärungen für NBTI gegeben werden. Im nächsten Abschnitt werden dann zwei Messmethoden vorgestellt, so, wie sie der JEDEC-Standard JESD90 [JED90] vorgibt, und eine Alternative, die Unzulänglichkeiten der ersten Methode zu umgehen versucht.

Bei Anliegen einer negativen Gate-Source-Spannung kann mit der Zeit eine Veränderung der Transistorparameter, vor allem der Schwellenspannung, beobachtet werden. Da PMOS-Transistoren in diesem Zustand betrieben werden, betrifft NBTI hauptsächlich PMOS-Transistoren. NMOS-Transistoren werden durch ein negatives Gatepotential kaum beeinflusst [BER06]. Die Auswirkung von positiven Gate-Source-Spannungen auf NMOS- oder PMOS-Transistoren, der so genannte PBTI-Effekt („Positive Bias Temperature Instability“), spielt nur bei Untersuchungen an neueren Technologien und Materialien eine Rolle [HEH10], [ROT12]. Alle weiteren Erklärungen beziehen sich auf NBTI-Effekte bei PMOS-Transistoren.

Bis vor einigen Jahren wurde NBTI überwiegend durch das so genannte „Reaktions-Diffusions-Modell“ (kurz: RD-Modell) erklärt [JEP77]. Dieses geht von einer elektrochemischen Reaktion zwischen Löchern und Defekten im Gateoxid an der Grenzfläche zum Siliziumsubstrat aus (engl. „interface state generation“)[14]. Bei einer zunehmenden Anzahl an positiv geladenen Defekten in der Nähe der Si-SiO_2-Grenzfläche ist dann zum Ausgleich am Gate eine negativere Spannung nötig, um den gleichen Stromfluss zu ermöglichen, d. h. der Betrag der Schwellenspannung der PMOS-Transistoren steigt. Die Degradation der Schwellenspannung erfährt aber nach Abschalten der Gate-Source-Spannung eine gewisse Ausheilung, d. h. der Betrag der Schwellenspannung nimmt wieder ab. Je höher die Temperatur und das vertikale elektrische Feld sind, desto stärker ist die Degradation. Die verschiedenen Varianten dieses Modells können aber viele der experimentell gewonnenen Daten nicht erklären. So geht das Modell beispielsweise davon aus, dass die Ausheilung der Degradation nach Abschalten des Stresses bei einer langen Stressdauer später einsetzt als bei

[14] Für genauere Erklärungen zu Haftstellen, festen Oxidladungen oder dem Einfang von Löchern siehe Kapitel 3.

einer kürzeren Stresszeit und die Ausheilung sich nur auf eine gewisse Zeit erstreckt. Dies widerspricht aber experimentellen Daten, in denen beobachtet wurde, dass unabhängig von der Stressdauer schon wenige Mikrosekunden nach Abschalten des Stresses eine Ausheilung stattfindet und diese logarithmisch mit der Zeit verläuft [REI06], [GRA11b]. Außerdem konnte festgestellt werden, dass eine positive Gate-Source-Spannung den Prozess der Ausheilung der Degradation beschleunigt, was gegen ein rein diffusionsbasiertes Modell spricht [GRA11].

Schon zu Beginn war aber auch die Idee vorhanden, dass das Einfangen von Löchern (engl. „hole trapping“) in prozessbedingte Defekte für die Parameteränderung verantwortlich sein könnte. Diese Idee wurde in den letzten Jahren weiterentwickelt [HUA06]. Neueste Ansätze gehen davon aus, dass diese „traps“ veränderlich sind. Wenn ein Defekt einmal erzeugt wurde, kann er je nach anliegendem Potential seine Ladung ändern [GRA09]. Dies könnte erklären, warum eine positive Gate-Source-Spannung die Ausheilung beschleunigt. Heute geht man davon aus, dass sowohl die „interface-state generation“ als auch das „hole trapping“ zur Parameterinstabilität beitragen. Die Degradation durch Löchereinfang ist dabei nur gering temperaturabhängig und am Anfang eines Stresszyklus dominant. Der Anteil der Grenzflächenhaftstellen-Generierung stimmt nach [MAH09] mit den Vorhersagen des RD-Modells überein. Dieser Anteil ist stark temperaturabhängig und folgt einem Potenzgesetz.

Eine vollständige Ausheilung der Parameterdegradation tritt im Normalfall nicht auf. Die Degradation wird deshalb in eine reversible und eine irreversible Komponente eingeteilt. Ob der reversible Anteil allein vom Löchereinfang und der irreversible Anteil durch die Haftstellengenerierung und -besetzung bestimmt werden, ist unklar [GRA11]. Die genaue Verbindung beider Mechanismen und ihre mögliche gegenseitige Abhängigkeit ist weiter Gegenstand der Forschung. Dies hängt auch davon ab, wie groß die Anzahl der schon vorher vorhandenen Defekte im Gateoxid ist und wie hoch das elektrische Feld und die Temperatur sind.

Neben der Erhöhung des Betrages der Schwellenspannung U_{th} kann als Folge von NBTI-Stress auch eine Reduzierung des Betrages des linearen Drain-Source-Stromes $I_{\mathrm{DS,lin}}$, des Sättigungs-Drain-Source-Stromes $I_{\mathrm{DS,sätt}}$ oder eine schwache Abnahme der Steilheit der Eingangskennlinie g_{m} stehen. Üblich ist aber die Untersuchung der Schwellenspannungsänderung.

NBTI ist ein stark technologieabhängiges Phänomen. Vor allem die Anwesenheit von Stickstoff kann sich negativ auswirken [STA06]. Fluor hingegen verringert die NBTI-Degradation, da es an der Grenzfläche von Silizium und Siliziumdioxid stabilere Bindungen eingeht als Wasserstoff [HOO01].

Messmethoden und Berechnung der Lebensdauer eines Transistors mit NBTI-Schädigung

Eine mögliche Vorgehensweise für die Untersuchung der Degradation der Transistorparameter durch NBTI-Effekte, im Englischen auch „Bias Temperature Stress (BTS)“ oder „High Temperature Gate Stress (HTGS)“ genannt, findet sich im JEDEC-Standard JESD90 [JED90].

1) Der PMOS-Transistor wird mit einer negativen Gate-Source-Spannung $U_{\mathrm{GS,stress}}$ gestresst, wobei $|U_{\mathrm{GS,max}}| < |U_{\mathrm{GS,stress}}| < |U_{\mathrm{BD}}/2|$ gilt. Dabei ist U_{BD} die Durchbruchspannung des

Gateoxids, also ungefähr 40 V für ein 40 nm dickes Oxid (vergleiche auch Kapitel 3). Source und Drain liegen dabei beide auf einem Potential von 0 V. Der Test wird bei maximaler Betriebstemperatur durchgeführt. Üblich sind dabei Temperaturen bis 200 °C [GRA11].

2) Die Spannungen werden für längere Zeit an den Transistor angelegt und die Änderung der Transistorparameter beobachtet. Für diese Beobachtung gibt es zwei Möglichkeiten mit jeweils einigen Vor- und Nachteilen.

2a) Die erste Methode ist im JEDEC-Standard JESD90 beschrieben. Weitere Informationen dazu finden sich im JEDEC-Dokument JP001. Wie im Fall der Hot Carrier - Messungen wird vor dem Anlegen des Stresses und dann in diskreten, logarithmisch äquidistanten Zeitabständen eine Eingangskennlinie aufgenommen. Zur Aufnahme der Kennlinien wird der Stress unterbrochen. Die Parameteränderung, vor allem die Änderung der Schwellenspannung, kann als Potenzgesetz auf eine festgelegte Änderung extrapoliert werden (siehe Abbildung 6.10).

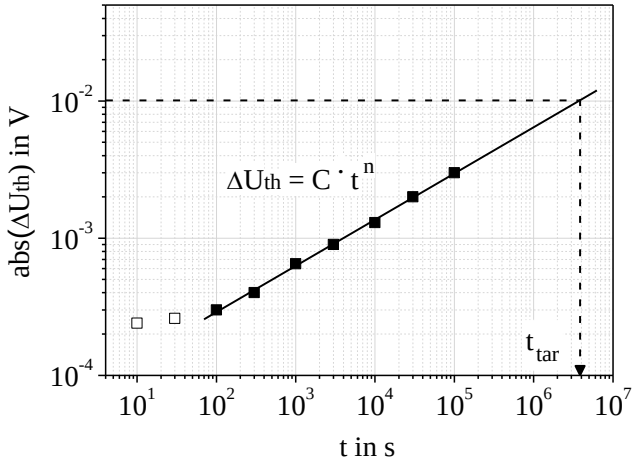


Abbildung 6.10: Beispiel für die betragsmäßige Änderung der Schwellenspannung in Abhängigkeit von der Zeit als Potenzgesetz; die Änderung wird auf den Messpunkt bei t = 0 s bezogen; mit einer Extrapolation lässt sich die Ausfallzeit t_{tar} ermitteln (hier bei abs(ΔU_{th}) = 10 mV), die in der Messzeit von 10.000 s nicht erreicht wurde

Typisch ist eine Akzeptanzgrenze von 10 mV Änderung der Schwellenspannung. Dieser Wert ist aber technologieabhängig und kann beispielsweise auch 50 mV betragen. Zur Zeitabhängigkeit der Schwellenspannungsänderung kommt gemäß dem JEDEC-Dokument JP001 eine Abhängigkeit von der Stressspannung hinzu:

$$\Delta U_{\mathrm{th}} \propto \exp\left(-\frac{D}{U_{\mathrm{GS,stress}}}\right) \qquad [6.22]$$

Dabei ist D eine Konstante. Berücksichtigt man auch die Temperaturabhängigkeit der Schwellenspannungsänderung, die in Anlehnung an Arrhenius Gleichung 6.23 folgt, erhält man eine kombinierte Abhängigkeit der Schwellenspannungsänderung von der Zeit, der Stressspannung und der Temperatur, aus der sich die Lebensdauer eines Transistors bei Betriebsbedingungen ermitteln lässt (Gleichung 6.24).

$$\Delta U_{\mathrm{th}} \propto \exp\left(-\frac{\Delta H}{k \cdot T}\right) \qquad [6.23]$$

$$\Delta U_{\mathrm{th}} \propto C \cdot t^{n} \cdot \exp\left(-\frac{D}{U_{\mathrm{GS,stress}}}\right) \cdot \exp\left(-\frac{\Delta H}{k \cdot T}\right) \qquad [6.24]$$

Hierbei ist $U_{\mathrm{GS,stress}}$ die Gate-Source-Stressspannung, C, n, D und ΔH sind Fit-Parameter[15], k ist die Boltzmann-Konstante und T die Temperatur in Kelvin. Zu beachten ist, dass C unabhängig von der Stressspannung ist, aber durchaus von der Temperatur abhängen kann. ΔH kann hingegen von der Stressspannung abhängig sein. Sowohl bei C als auch bei ΔH ist zudem eine Abhängigkeit von der Zeit möglich. Die Konstante n ist unabhängig von der Zeit, aber abhängig von der Stressspannung und der Temperatur.

Der Nachteil dieser Methode ist die Unterbrechung des Stresses während der Messung der Eingangskennlinie. In der Zeit zwischen dem Ende eines Stressintervalls und dem Beginn des nächsten Stressintervalls kann je nach Dauer des Stresses schon nach wenigen Sekunden Pause ein beachtlicher Anteil der Degradation wieder ausgeheilt sein. Deshalb sollte diese Methode nur angewendet werden, wenn die Schwellenspannung sehr schnell nach Abschalten des Stresses gemessen werden kann, oder wenn die Möglichkeit vorhanden ist, den Anteil der Ausheilung genau zu bestimmen.

2b) Da es schwierig ist, die Zeit zwischen Stress und Messung klein zu halten, gibt es eine alternative Methode, bei der während des Stresses gleichzeitig gemessen wird. Dazu muss eine kleine Drain-Source-Spannung U_{DS} angelegt und der Drain-Source-Strom I_{DS} simultan gemessen werden [STA06]. Der große Vorteil dieser Methode ist, dass keine Ausheilung der Degradation stattfinden kann, da der Stress nie unterbrochen wird. Es ist also möglich, die komplette NBTI-Degradation zu detektieren. In der Literatur findet man diese Methode unter dem Begriff „On-the-fly“-Charakterisierung [DEN04]. Der Nachteil der Methode ist die erschwerte Bestimmung der Schwellenspannungsänderung. Oft wird in regelmäßigen Abständen ein kleiner Gatepuls um $U_{\mathrm{GS,stress}}$ herum angelegt, was die Bestimmung von g_{m} und damit der Schwellenspannung ermöglicht. Wenn es aber nicht möglich ist, g_{m} zu bestimmen, geht man für die Berechnung der Änderung der Schwellenspannung ΔU_{th} zum Zeitpunkt t in erster Näherung davon aus, dass sich die Ladungsträgerbeweglichkeit nicht ändert [MAH09], [HER11]:

$$\Delta U_{\mathrm{th}}(t) = \frac{\Delta I_{\mathrm{DS}}(t)}{I_{\mathrm{DS}}(0)} \cdot \left(U_{\mathrm{GS,stress}} - U_{\mathrm{th}}(0)\right) \qquad [6.25]$$

Hierbei ist $\Delta I_{\mathrm{DS}}(t)$ die Änderung des Drain-Source-Stromes nach der Zeit t im Vergleich zu der Zeit $t = 0$ s, wo $I_{\mathrm{DS}}(0)$ vorliegt, $U_{\mathrm{GS,stress}}$ bezeichnet die Drain-Source-Stressspannung und $U_{\mathrm{th}}(0)$ ist die Schwellenspannung des Transistors vor dem Stress.

Untersuchungen haben aber gezeigt, dass die Änderung des linearen Drainstromes nicht ausschließlich durch die Schwellenspannungsänderung hervorgerufen wird, sondern zu einem gewissen Anteil auch von der Änderung der Ladungsträgerbeweglichkeit abhängt [HER11], [OTA12]. Damit ist eine genaue Berechnung der Schwellenspannungsänderung nur möglich,

[15] Man beachte, dass ΔH hier keine Aktivierungsenergie im Sinne von Arrhenius darstellt, da nicht die Temperaturabhängigkeit der Reaktionszeit bzw. der Reaktionsgeschwindigkeit betrachtet wird, sondern die Temperaturabhängigkeit der Schwellenspannung.

wenn der Drain-Source-Strom bei verschiedenen Gate-Source-Spannungen gemessen und g_m bestimmt wird. Wird der Drain-Source-Strom allerdings immer nur bei einer festen Gate-Source-Spannung, nämlich der Stressspannung, gemessen, ist es sehr schwierig, eine Schwellenspannung zu berechnen.

In der Literatur finden sich noch weitere Ansätze für NBTI-Untersuchungen. Beispielsweise können über spezielle Schaltungen die Auswirkungen von NBTI gemessen werden, ohne einen Teil der Degradation durch Ausheilung zu verlieren. So kann aus der zeitlichen Frequenzänderung eines Ringoszillators in Betrieb auf eine Änderung der Schwellenspannung geschlossen werden [KIM08], [KIM11]. Es ist nur darauf zu achten, dass dann in der Schaltung andere Zuverlässigkeitsprobleme, wie z.B. die Entstehung von „heißen“ Ladungsträgern, vermieden werden.

In den im folgenden Abschnitt vorgestellten Untersuchungen wurde zunächst die Methode des JEDEC-Standards (2a) angewandt. Da eine Änderung der Schwellenspannung aber nicht eindeutig erkennbar war, wurde anschließend die zweite Methode getestet (2b). Dabei wurde der Drain-Source-Strom immer nur bei einer Gate-Source-Spannung (der Stressspannung) gemessen. Statt zu versuchen, aus den Messungen die Schwellenspannungsänderung zu berechnen, wurde untersucht, welche Gesetzmäßigkeiten für die Änderung des Drain-Source-Stromes gelten. Würde man dann eine Akzeptanzgrenze für die Änderung des Stromes festlegen, könnte man die Lebensdauer eines Transistors berechnen, ohne auf die Messung (oder Berechnung) der Schwellenspannungsänderung angewiesen zu sein.

NBTI-Effekte beim H10-Prozess: JEDEC-Messmethode vs. Strom-Messung

Mit dem verwendeten Messaufbau war es nicht möglich, die Eingangskennlinien in weniger als einer Minute zu messen. Abbildung 6.11 zeigt die Ergebnisse einer Stressmessung nach der ersten Methode (2a). Die Schwellenspannungen wurden mit Hilfe der Tangentenmethode bestimmt. Auf der y- Achse sind die Absolutwerte der Änderungen aufgetragen, um sie logarithmisch darstellen zu können. Die eigentliche Änderung ist negativ, d. h. die Schwelle wird wegen der zunehmenden Anzahl an positiv geladenen Defekten im Gateoxid negativer.

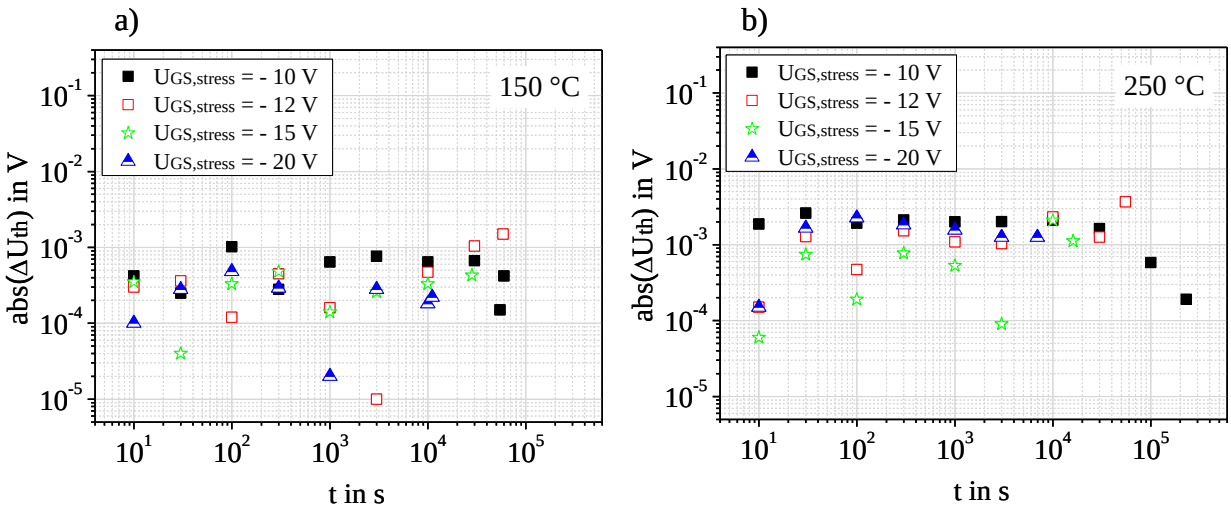


Abbildung 6.11: Betrag der Änderung der Schwellenspannung bezogen auf den Messwert $U_{th,0}$ nach t = 0 s in Abhängigkeit von der Zeit für vier verschiedene Gate-Source-Stressspannungen $U_{GS,stress}$; die Potentiale an Drain, Source und Back-Gate betragen jeweils 0 V; Messungen an PMOS-Transistoren der Länge l = 1,6 µm und der Weite w = 36 µm bei 150 °C (a) und bei 250 °C (b)

Im Rahmen der Messgenauigkeiten sind bei beiden Temperaturen keine signifikanten Änderungen feststellbar, vor allem aber ist keine Tendenz im zeitlichen Verlauf sichtbar. Deshalb ist es nicht möglich, aus den gewonnenen Daten korrekt auf die Zeit zu extrapolieren, nach der eine Schwellenspannungsänderung von zum Beispiel 50 mV erreicht wäre. Die erkennbaren Schwankungen der Schwellenspannungen sind auf die Auflösungsgrenze des Messgerätes zurückzuführen. Trotz dieser Schwankung müsste eine Änderung der Schwellenspannung aber sichtbar sein, wenn die Ausheilung durch die Stresspausen während der Eingangskennlinienmessung nicht zu groß wäre, um den NBTI-Effekt messen zu können. Aus diesem Grund wurden die weiteren Messungen nach der zweiten Methode (2b) durchgeführt. Dabei wurde der Drain-Source-Strom in äquidistanten Zeitabständen gemessen während weiterhin die Gate-Source-Stressspannung anlag. In Abbildung 6.12 sind die Ergebnisse dieser Messungen bei verschiedenen Temperaturen und Gate-Source-Stress-spannungen als Änderungen des Drain-Source-Stromes mit der Zeit dargestellt. Dabei war der Drain-Source-Strom selbst negativ und einige Mikroampère groß, abhängig von der Stress-spannung und der Temperatur. U_{DS} betrug -0,05 V, am Back-Gate lag jeweils ein Potential von 0 V an und die Stressdauer war immer mindestens 100.000 s lang.

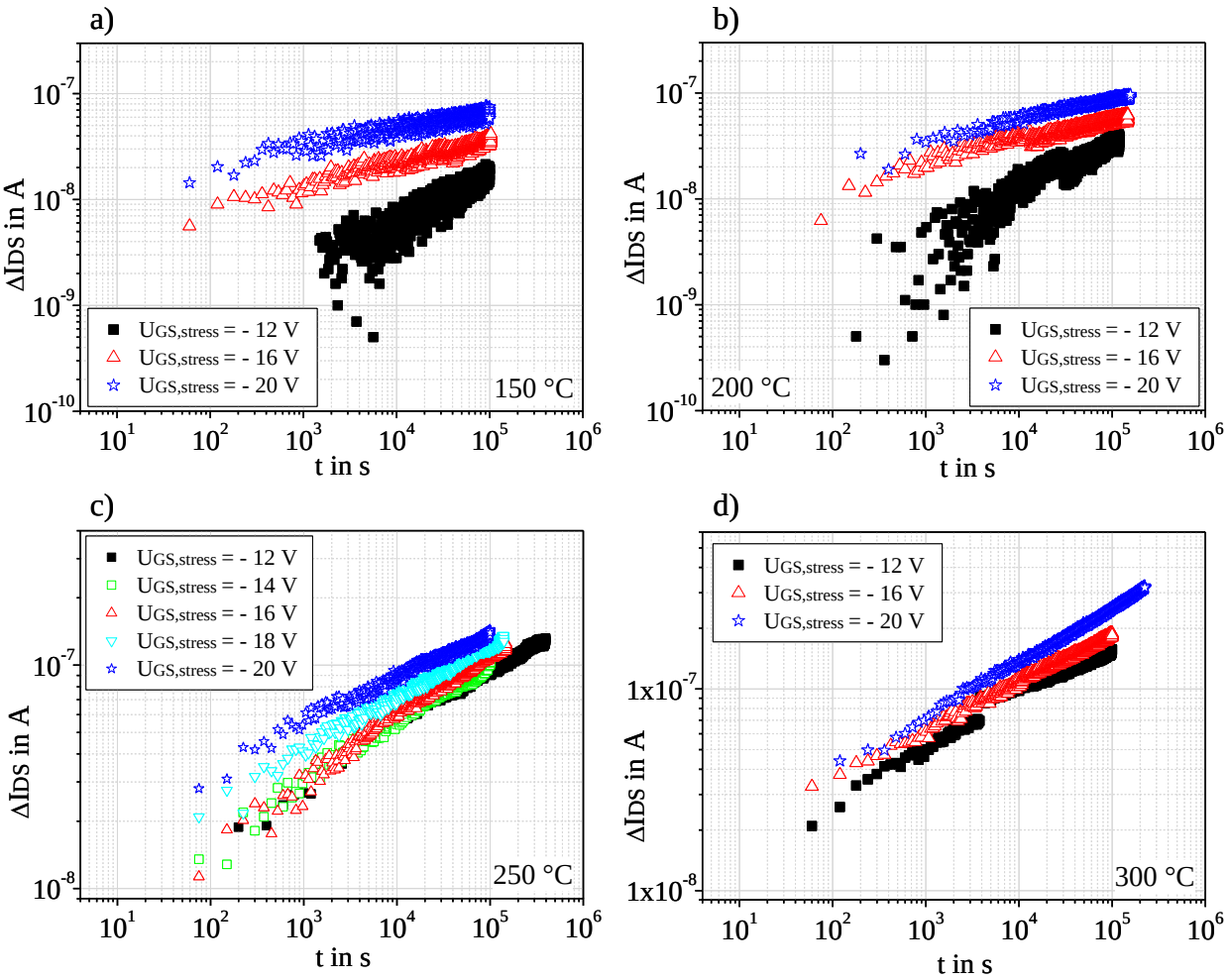


Abbildung 6.12: Änderung des Drain-Source-Stromes I_{DS} bezogen auf den Messwert nach t = 0 s in Abhängigkeit von der Zeit bei $U_{GS,stress}$ = -12 V bis -20 V; U_{DS} = -50 mV und U_{BGS} = 0 V; Messungen an PMOS-Transistoren der Länge l = 1,6 µm und Weite w = 36 µm bei 150 °C (a), 200 °C (b), 250 °C (c) und 300 °C (d)

Der Absolutwert des Drain-Source-Stromes I_{DS} nimmt mit der Zeit ab. Die Breite der Kurven in Abbildung 6.12 spiegelt die Temperaturschwankung des Thermochucks wider. Bei der kleinsten gewählten Stressbedingung (T = 150 °C und $U_{GS,stress}$ = -12 V) erreicht man in 100.000 s eine Stromänderung von ΔI_{DS} = 1,53 · 10^{-8} A, bei der größten Stressbedingung (T = 300 °C und $U_{GS,stress}$ = -20 V) kann man in dieser Zeit eine etwa 17-mal so große Änderung von ΔI_{DS} = 2,54 · 10^{-7} A verzeichnen. Das Stromniveau zu Beginn der Messungen liegt je nach Temperatur und Stressspannung zwischen 4 · 10^{-6} A und etwa 8 · 10^{-6} A.

Abhängigkeit der Drain-Source-Stromänderung von der Zeit, der Stressspannung und der Temperatur

Die zeitlichen Änderungen der gemessenen Drain-Source-Stromdifferenz folgen einem Potenzgesetz. Analog zur Zeitabhängigkeit der Schwellenspannung (vergleiche Abbildung 6.10) gilt für die Änderung des Drain-Source-Stromes:

$$\Delta I_{DS} \propto C \cdot t^{n} \qquad [6.26]$$

Um die Abhängigkeit der Stromänderung von der angelegten Stressspannung zu bestimmen, wurde in Anlehnung an Gleichung 6.22 der Logarithmus der Stromänderung über $1/U_{GS,stress}$ aufgetragen. Abbildung 6.13 stellt diese Abhängigkeit dar. ΔI_{DS} ist dabei die Differenz zwischen dem Messwert des Stromes nach t = 0 s und dem Messwert nach t = 50.000 s bzw. t = 100.000 s. Die Differenz zwischen t = 0 s und t = 100.000 s wurde gewählt, um eine möglichst große Stromänderung in die weiteren Berechnungen mit einbeziehen zu können. Die Differenz zwischen t = 0 s und t = 50.000 s dient als Vergleich dazu.

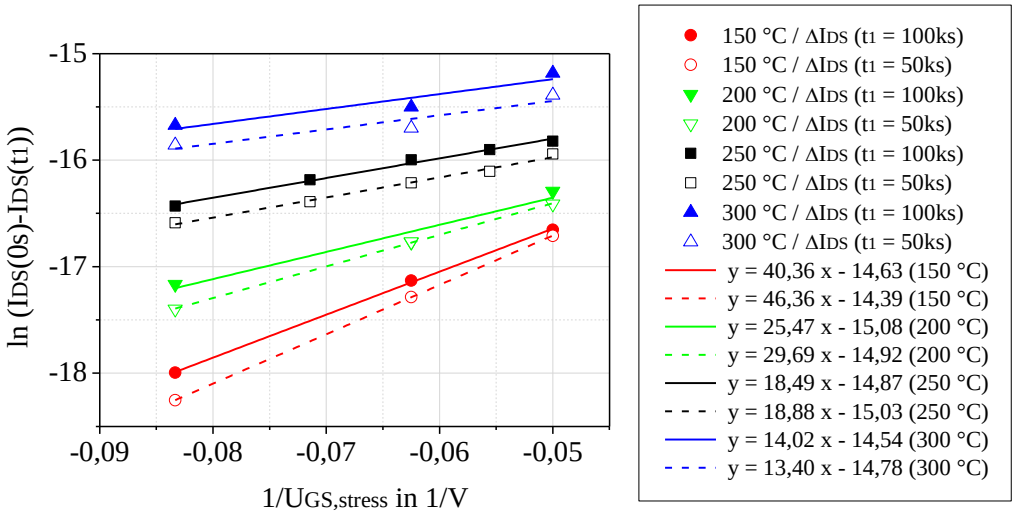


Abbildung 6.13: Logarithmus der Differenz des Drain-Source-Stromes vor dem Stress (t = 0 s) und nach 100.000 s (ausgefüllte Symbole) bzw. 50.000 s Stress (leere Symbole) in Abhängigkeit von $1/U_{GS,stress}$; Messungen an PMOS-Transistoren der Länge l = 1,6 µm und der Weite w = 36 µm bei 150 °C, 200 °C, 250 °C und 300 °C; t_1 ist entweder t_1 = 50.000 s oder t_1 = 100.000 s (siehe Legende)

Wie Abbildung 6.13 zeigt, gilt für die Änderung des Drain-Source-Stromes analog zu Gleichung 6.22 folgender Zusammenhang, wobei D eine temperaturabhängige Konstante ist:

$$\Delta I_{\mathrm{DS}} \propto \exp\left(-\frac{D(T)}{U_{\mathrm{GS,stress}}}\right) \qquad [6.27]$$

Die dargestellte Abhängigkeit wird für höhere Temperaturen kleiner. Die Konstante $D(T)$ halbiert sich ungefähr alle 100 °C. Gleichung 6.27 gilt sowohl für die Differenz der Stromwerte zwischen t = 0 s und t = 50.000 s als auch für die Differenz zwischen t = 0 s und t = 100.000 s. Eine Zeitabhängigkeit von D lässt sich durch Abbildung 6.13 weder bestätigen noch ausschließen. Die kleinen Differenzen in den Steigungen der durchgezogenen und der gestrichelten Geraden könnten alleine durch die Schwankung der Stromwerte (vergleiche Abbildung 6.12) erklärt werden.

Aus den Ergebnissen von Abbildung 6.13 kann in Anlehnung an Gleichung 6.23 für die einzelnen Stressspannungen die Temperaturabhängigkeit der Stromänderung aufgetragen und die jeweilige Konstante ΔH bestimmt werden.

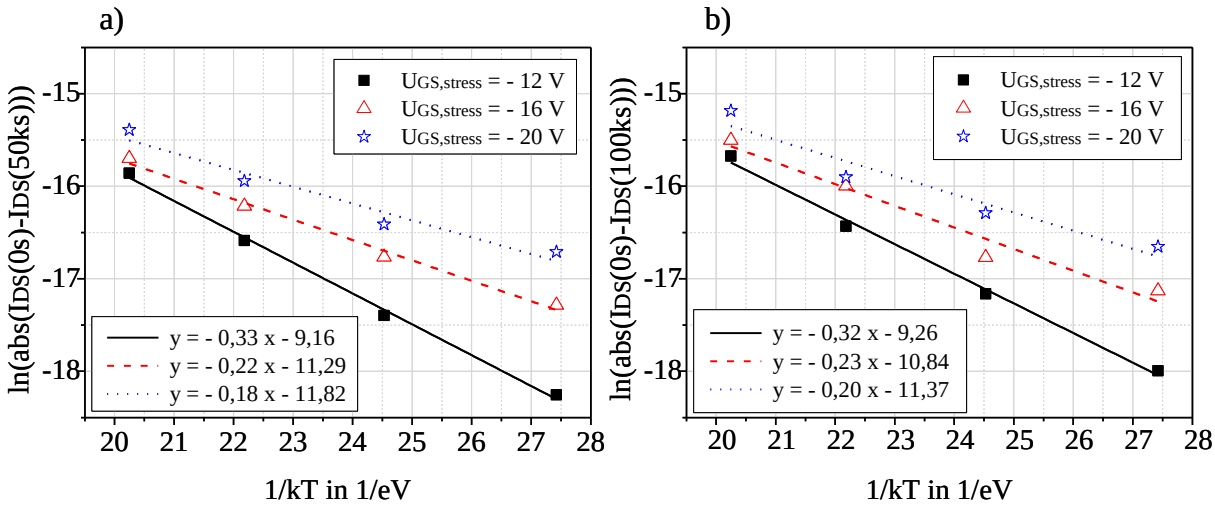


Abbildung 6.14: Logarithmus der Differenz des Drain-Source-Stromes vor dem Stress (t = 0 s) und nach 50.000 s (a) bzw. 100.000 s (b) in Abhängigkeit von 1/kT; Messungen an PMOS-Transistoren der Länge l = 1,6 µm und der Weite w = 36 µm bei Gate-Source-Stressspannungen von -12 V, -16 V und -20 V

Abbildung 6.14 zeigt, dass der Zusammenhang aus Gleichung 6.23 auch für die Änderung des Stromes gilt. ΔH ändert wird mit zunehmender Stressspannung größer und erreicht Werte zwischen 0,18 eV und 0,33 eV. Diese Ergebnisse passen gut zu dem im JEDEC-Dokument JP001 angegebenen Bereich von 0,2 eV bis 0,6 eV. Ob ΔH auch von der Zeit abhängt, lässt sich mit den bisherigen Ergebnissen schwer sagen. Die Differenzen der Steigungen in Abbildung 6.14 a) und b) lassen sich auch leicht auf die Schwankung der Stromwerte (vergleiche Abbildung 6.12) zurückführen. Es gilt somit analog zu Gleichung 6.23:

$$\Delta I_{\mathrm{DS}} \propto \exp\left(-\frac{\Delta H(U_{\mathrm{GS,stress}})}{k \cdot T}\right) \qquad [6.28]$$

Die in den Gleichungen 6.26 bis 6.28 dargestellten Zusammenhänge zwischen der Änderung des Drain-Source-Stromes und der Zeit, der Gate-Source-Stressspannung und der Temperatur gleichen den Gesetzmäßigkeiten für die Abhängigkeit der Schwellenspannungsänderung von

ebendiesen Parametern (Gleichungen 6.22 bis 6.24). Dies bedeutet entweder, dass die Degradation von g_m, d. h. der Ladungsträgerbeweglichkeit, keine entscheidende Rolle spielt, oder, dass auch die Änderung von g_m den dargestellten Gleichungen folgt. In jedem Fall aber ist es möglich, Degradationen zu beobachten, ohne die Schwellenspannung explizit zu messen.

Zudem ist es auch nicht nötig, die Schwellenspannungsänderung zu berechnen was die hier vorgestellte Methode von üblichen „On-the-fly"-Charakterisierungen unterscheidet. Man müsste nur eine maximal erlaubte Änderung des Drain-Source-Stromes festlegen, die Transistoren nicht überschreiten dürfen, wenn sie als zuverlässig angesehen werden sollen. Aus Abbildung 6.12 kann man dazu entnehmen, dass innerhalb von 10 Jahren bei der maximal erlaubten Betriebsspannung von $U_{GS,max}$ = -12 V und der maximalen Betriebstemperatur von 250 °C eine Änderung des Drain-Source-Stromes von weniger als $0{,}25 \cdot 10^{-6}$ A zu erwarten wäre. Bei einem Stromniveau von etwa $5 \cdot 10^{-6}$ A zu Beginn der Messung entspricht dies einer Stromänderung von 5 %, was im Rahmen der üblichen akzeptierten Änderungen (z.B. bei den Hot Carrier - Messungen) liegt.

Verlauf der Ausheilung der Stromdegradation

Um zu bestätigen, dass es sich bei den beobachteten Stromdegradationen tatsächlich um einen NBTI-Effekt handelt, wird nun im Folgenden gezeigt, dass sich die Degradationen ausheilen lassen, so wie es bei NBTI erwartet werden würde. Dazu wurde ein PMOS-Transistor 7000 s lang und ein anderer 70.000 s lang mit $U_{GS,stress}$ = -12 V gestresst und danach jeweils über die gleiche Zeitdauer ohne Spannungsbelastung am Gate beobachtet. Der Drain-Source-Strom wurde während der Stresszeit in regelmäßigen Abständen bei $U_{GS,stress}$ = -12 V gemessen. In der Ausheilungsphase lag für die Messungen des Drain-Source-Stromes ebenfalls U_{GS} = -12 V für wenige Sekunden am Gate an. Während der Stresspause waren alle Anschlüsse floatend, in der Stress- und Messphase lagen 0 V an Source und Back-Gate sowie -0,05 V am Drain an. Die Temperatur des Chucks betrug während der kompletten Stress- und Ausheilungszeit 250 °C. In Abbildung 6.15 a) ist der Verlauf des Drain-Source-Stromes für die beiden Transistoren dargestellt. Abbildung 6.15 b) zeigt den Verlauf der Ausheilung. Die horizontalen Linien markieren das Niveau von 50 % des Stromwertes am Ende der Stresszeit.

Es zeigt sich, dass in beiden Fällen nach einer Stresspausenzeit von der Dauer des vorherigen Stresses ungefähr die Hälfte der Degradation rückgängig gemacht werden kann, wenn die Stresstemperatur und die Pausentemperatur 250 °C betragen. Dies stimmt mit den Beobachtungen des RD-Modells bezüglich der Schwellenspannungsänderung überein, das eine Halbierung der Degradation nach einer Ausheilungszeit von der Dauer des Stresses vorhersagt wird [laut GRA11]. Andererseits ist der zeitliche Verlauf der Ausheilung aber logarithmisch und bestätigt damit auch neuere Erkenntnisse zur Ausheilung von NBTI [GRA11]. Wie schnell die Ausheilung nach Abschalten des Stresses einsetzt, ist mit den gegebenen Messbedingungen schwer zu überprüfen.

Wenn der Stress bei 250 °C erfolgt, der Thermochuck aber während der Stresspause auf 300 °C geheizt wird, kann nach einer Stresspause derselben Länge wie der vorherigen Stressdauer die Degradation fast komplett ausgeheilt werden. Abbildung 6.16 demonstriert dieses Ergebnis für eine Stress- und Pausendauer von 7000 s. Die Messungen des Stromes während des Stresses und nach der Pause wurden bei 250 °C durchgeführt. Auch dies stimmt mit Beobachtungen in der Literatur für den Verlauf der Ausheilung der Schwellenspannungsänderung überein [KAT08], [HUA10].

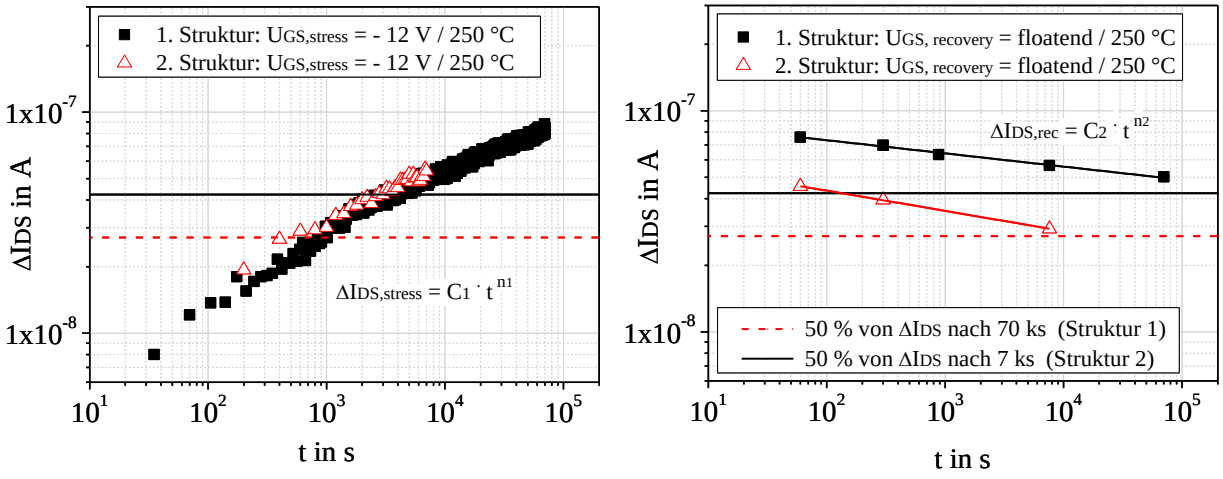


Abbildung 6.15: Änderung des Drain-Source-Stromes I_{DS} bezogen auf den Messwert nach t = 0 s in Abhängigkeit von der Zeit bei $U_{GS,stress}$ = -12 V; U_{DS} = -0,05 V und U_{BGS} = 0 V während des Stresses (a) bzw. $U_G = U_D = U_S = U_{BG}$ = floatend während der Ausheilungsphase (b); Messungen an PMOS-Transistoren der Länge l = 1,6 µm und der Weite w = 36 µm; T = 250 °C; Stress- und Gesamtpausendauer = 70.000 s (ausgefüllte Kästchen) bzw. 7000 s (leere Dreiecke)

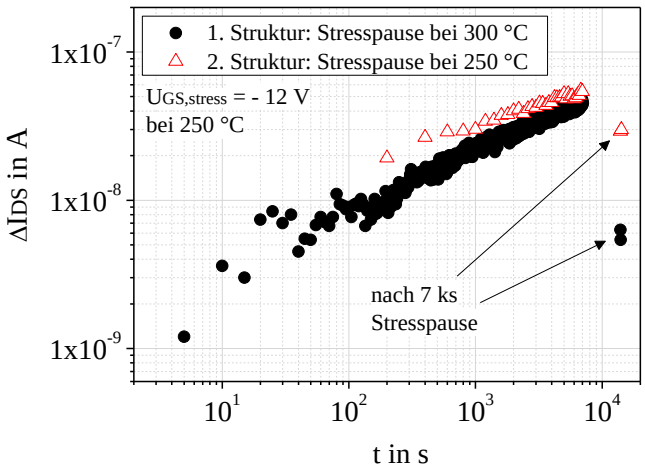


Abbildung 6.16: Änderung des Drain-Source-Stromes I_{DS} bezogen auf den Messwert nach t = 0 s in Abhängigkeit von der Zeit bei $U_{GS,stress}$ = -12 V; $U_{DS,stress}$ = -0,05 V und $U_{BGS,stress}$ = 0 V; Messungen an PMOS-Transistoren der Länge l = 1,6 µm und der Weite w = 36 µm; T = 250 °C während des Stresses und während der Messungen vor und nach der Pause; T = 300 °C während der Stresspause des 1. Transistors (ausgefüllte Kreise) und T = 250 °C während der Stresspause des 2. Transistors (leere Dreiecke)

Abbildungen 6.15 und 6.16 zeigen, dass der Verlauf der Ausheilung der Stromdegradation mit den Angaben in der Literatur für den Verlauf der Ausheilung der Schwellenspannungsänderung übereinstimmt. Die Defektausheilung ist dabei eine Reaktion, die thermisch beschleunigt werden kann (fast komplette Ausheilung bei 300 °C) und zudem von der Anzahl der vorhandenen Defekte abhängt (logarithmischer Zeitverlauf). Damit konnte gezeigt werden, dass es sich bei den Stromdegradationen tatsächlich um einen NBTI-Effekt handelt.

Fazit

Zwei wichtige Erkenntnisse konnten aus den Messungen zur Parameterinstabilität bei negativer Gate-Source-Spannung gewonnen werden. Zum einen ist es möglich, zur Untersuchung von NBTI-Effekten statt der Änderung der Schwellenspannung die Änderung des Drain-Source-Stromes zu betrachten, denn beide Änderungen folgen den gleichen Gesetzen bezüglich ihrer Abhängigkeit von der Zeit, der Stressspannung, der Temperatur und der Ausheilung. Zudem sind die Gesetzmäßigkeiten nicht nur bis 200 °C, sondern auch bis 300 °C gültig. Da die Messung des Stromes bei der Gate-Source-Stressspannung erfolgt, wird damit das Problem der zwischenzeitlichen Ausheilung, das bei der Stressunterbrechung zur Bestimmung der Schwellenspannung vorliegt, umgangen. Sofern man nicht explizit auf die genaue Kenntnis der Schwellenspannungsänderung angewiesen ist, stellt die Beobachtung der Stromdegradation eine einfach zu realisierende Alternative dar. Es ist lediglich nötig, eine maximale Stromänderung festzulegen, die nicht überschritten werden darf.

6.2 Ringoszillatoren als digitale Grundschaltungen

6.2.1 Theoretische Aspekte zu Ringoszillatoren

Ein Ringoszillator ist eine einfache digitale Grundschaltung, die als Taktgeber für andere Schaltungen dienen kann. Er besteht aus einer Kette von Invertern. Jeder dieser CMOS-Inverter ist aus einem PMOS- und einem NMOS-Transistor aufgebaut (siehe Abbildung 6.17), von denen immer einer leitet und der andere sperrt. Wird die Versorgungsspannung U_{dd} auf den Eingang gegeben, sperrt der PMOS-Transistor wegen $U_{GS} = U_1 - U_{dd} = 0$ V, der NMOS-Transistor aber leitet mit $U_{GS} = U_1 = U_{dd}$. Das Ausgangssignal ist mit der Masse verbunden und entspricht $U_2 = 0$ V. Liegen am Eingang hingegen 0 V an, so sperrt der NMOS-Transistor wegen $U_{GS} = U_1 = 0$ V, der PMOS-Transistor leitet aber mit $U_{GS} = U_1 - U_{dd} = -U_{dd}$. Das Ausgangssignal ist in dieser Konfiguration mit der Versorgungsspannung U_{dd} verbunden und wiederum genau invertiert zum Eingangssignal. Der Inverter hat also zwei logische Zustände, „1“ (U_{dd}) und „0“ (0 V).

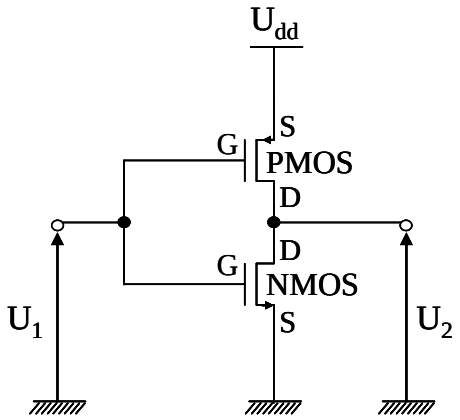


Abbildung 6.17: Schema eines CMOS-Inverters, aus dem die Ringoszillatorschaltung aufgebaut ist; U_{dd} ist die Versorgungsspannung, U_1 die Eingangsspannung und U_2 die Ausgangsspannung

In einem Ringoszillator sind mehrere Inverter hintereinander geschaltet. Das Ausgangssignal des Inverters n entspricht damit dem Eingangssignal des Inverters n+1. Sind die Inverter in einem geschlossenen Schaltkreis verbunden, bildet das Ausgangssignal des letzten Inverters das Eingangssignal des ersten Inverters. Bei einer ungeraden Anzahl an Invertern ändert sich nach jedem Durchlauf das Eingangssignal des ersten und damit jedes folgenden Inverters: der Schaltkreis oszilliert. Wenn das ursprüngliche Signal U_1 einmal vorgegeben wurde, oszilliert der Ringoszillator von alleine, wenn weiterhin die Versorgungsspannung U_{dd} sowie das Massepotential anliegen.

Abbildung 6.18 zeigt den schematischen Aufbau eines solchen Ringoszillators, das auch den in dieser Arbeit untersuchen Ringoszillatoren entspricht. Der Buffer dient dazu, das Ausgangssignal zu entkoppeln und so separat messen zu können.

Die Frequenz f, mit der der Ringoszillator schwingt, lautet [GRA99]:

$$f = \frac{1}{n \cdot (t_{d,n} + t_{d,p})} \qquad [6.29]$$

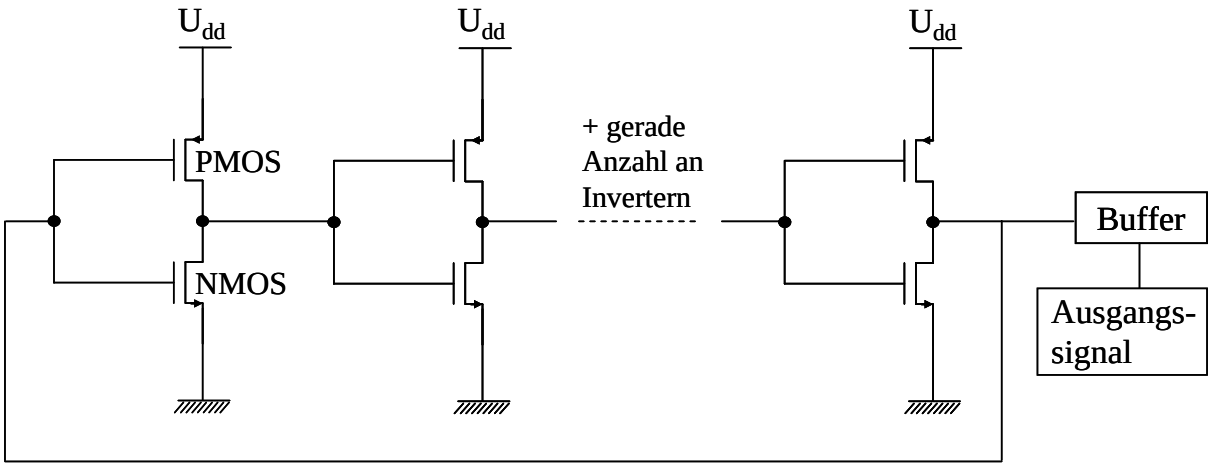


Abbildung 6.18: Schema eines Ringoszillators mit einer ungeraden Anzahl an Invertern und einem Buffer zum Abgreifen des Ausgangssignals; U_{dd} ist die Versorgungsspannung

Hierbei ist n die Anzahl der Inverter und $t_{d,n}$ sowie $t_{d,p}$ die Verzögerungszeiten des Signals pro Inverter (siehe nachfolgenden Text). Diese Verzögerungszeiten hängen von der Geometrie der Transistoren und deren Schwellenspannungen, der Gateoxidkapazität, der Versorgungsspannung, der Ladungsträgermobilitäten sowie der Kapazität des Ausgangs ab.

Die Verzögerungszeit $t_{d,n}$ im Falle eines leitenden NMOS-Transistors, d.h. wenn das Signal von „1“ zu „0“ wechselt, wird von der Geometrie und der Schwellenspannung des NMOS-Transistors $U_{th,n}$ sowie von der Beweglichkeit der Elektronen μ_n bestimmt [WES85]:

$$t_{d,n} = \frac{C_L \cdot l_n}{\mu_n \cdot C_{ox} \cdot w_n \cdot (U_{dd} - U_{th,n})} \cdot \left[\frac{2 \cdot (U_{th,n} - 0{,}1 \cdot U_{dd})}{U_{dd} - U_{th,n}} + \ln \frac{1{,}9 \cdot U_{dd} - 2 \cdot U_{th,n}}{0{,}1 \cdot U_{dd}} \right] \quad [6.30]$$

Wenn der PMOS-Transistor leitet, wechselt das Signal von „0“ zu „1“ und die Verzögerungszeit $t_{d,p}$ ist von der Geometrie und der Schwellenspannung des PMOS-Transistors $U_{th,p}$ sowie von der Beweglichkeit der Löcher μ_p anhängig [WES85]:

$$t_{d,p} = \frac{C_L \cdot l_p}{\mu_p \cdot C_{ox} \cdot w_p \cdot (U_{dd} - |U_{th,p}|)} \cdot \left[\frac{2 \cdot (|U_{th,p}| - 0{,}1 \cdot U_{dd})}{U_{dd} - |U_{th,p}|} + \ln \frac{1{,}9 \cdot U_{dd} - 2 \cdot |U_{th,p}|}{0{,}1 \cdot U_{dd}} \right] \quad [6.31]$$

In den Gleichungen 6.30 und 6.31 ist C_L die Ausgangskapazität eines Inverters, l_n und l_p sind die Kanallängen und w_n und w_p die Kanalweiten von NMOS- und PMOS-Transistor, C_{ox} ist die Gateoxidkapazität und U_{dd} die Versorgungsspannung des Ringoszillators. Zusätzliche Effekte wie die Zeit zur Ausbildung des Kanals oder unterschiedliche Signalformen haben keinen wesentlichen Einfluss auf die Verzögerungszeit und wurden deshalb hier vernachlässigt.

Die Ladungsträgermobilitäten $\mu_n = \mu_n(T)$ und $\mu_p = \mu_p(T)$ sowie die Schwellenspannungen $U_{th,n} = U_{th,n}(T)$ und $U_{th,p} = U_{th,p}(T)$ sind temperaturabhängig. Daraus ergibt sich für die Temperaturabhängigkeit der Ringoszillatorfrequenz:

$$f = f\left(U_{th,n}(T), U_{th,p}(T), \mu_n(T), \mu_p(T)\right) \qquad [6.32]$$

Um eine gleichmäßige Oszillation des Ringoszillators (symmetrische Signalflanken) zu gewährleisten, sollten die Verzögerungszeiten $t_{d,n}$ und $t_{d,p}$ gleich sein. Wenn die Beträge der Schwellenspannungen gleich sind, kann der Unterschied in der Mobilität der Ladungsträger ausgeglichen werden, indem PMOS-Transistoren eingesetzt werden, deren Kanalweiten w_p um den Faktor μ_n/μ_p größer sind als die Kanalweiten w_n der verwendeten NMOS-Transistoren. Es ist aber auch möglich, eine andere Kombination von $U_{th,n}$, $U_{th,p}$, w_n und w_p zu wählen, um die Verzögerungszeiten anzugleichen.

Die Untersuchungen in Abschnitt 6.2.2 konzentrieren sich auf zwei Aspekte: die Abhängigkeit der Frequenz von der Temperatur sowie die Langzeitstabilität der Frequenz bei hoher Temperatur. Für den ersten Punkt wurden Ringoszillatoren von Raumtemperatur bis 450 °C charakterisiert, für den zweiten Punkt lagerten sie in Betrieb in Öfen bei 250 °C und 350 °C. Neben der Frequenz wurde auch die Veränderung der Stromaufnahme I mit der Temperatur bzw. mit der Zeit beobachtet.

6.2.2 Charakterisierung und Langzeitverhalten von Ringoszillatoren

Die untersuchten Ringoszillatoren sind aus 101 CMOS-Invertern zusammengesetzt. Die Kanallängen der NMOS- und PMOS-Transistoren betragen jeweils 1,6 µm, die Kanalweiten sind mit w_n = 3,6 µm und w_p = 7,2 µm um einen Faktor zwei verschieden. Bei Raumtemperatur beträgt die Schwellenspannung des verwendeten NMOS-Transistors ungefähr $U_{th,n}$ = 1 V, die Schwellenspannung des PMOS-Transistors circa $U_{th,p}$ = -2 V. Die verwendeten NMOS- und PMOS-Transistoren entsprechen den im Abschnitt 6.1.2 charakterisierten Transistoren. Die Versorgungsspannung von Ringoszillator und Buffer beträgt jeweils 5 V.

Alle Charakterisierungen und Langzeituntersuchungen der Ringoszillatoren und auch der Bandgap-Referenzen wurden nicht auf Waferebene sondern an aufgebauten Schaltungen durchgeführt, die in den beiden Öfen geheizt wurden. Für weitere Informationen zum hochtemperaturtauglichen Aufbau von Chips in Keramikgehäusen siehe auch Kapitel 2.

Temperaturabhängigkeit von Frequenz und Stromaufnahme

In Abbildung 6.19 ist die Frequenz des Ringoszillators in Abhängigkeit von der Temperatur dargestellt (siehe auch [GRE12]). Die schwarzen Kästchen zeigen die gemessenen Werte, die beiden Linien die mit der Software Cadence® erstellten Simulationen. Dabei bezeichnet C1 die Simulation der Frequenz, wenn von den betragsmäßig minimal erlaubten Schwellenspannungen von NMOS- und PMOS-Transistor ($\pm$ |U_{th} − 0,1 V|) ausgegangen wird und C4 den Fall für die betragsmäßig maximal erlaubten Schwellenspannungen ($\pm$ |U_{th} + 0,1 V|). Die minimalen bzw. maximalen Schwellenspannungen bilden die erlaubten technologischen Grenzen. In der Simulation ist die Temperaturabhängigkeit der Schwellenspannung und der Ladungsträgermobilität berücksichtigt.

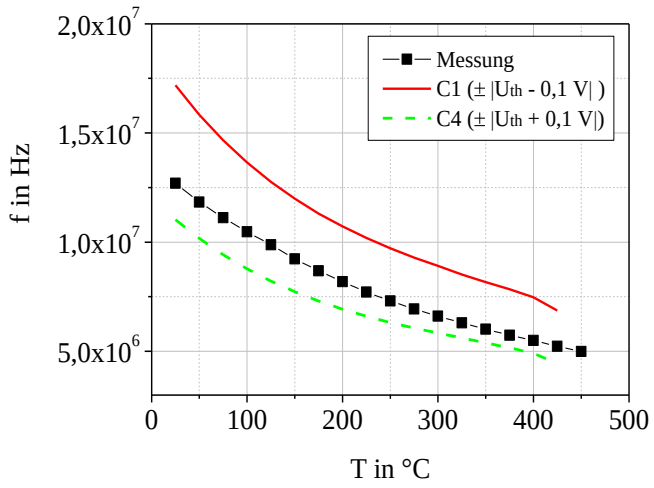


Abbildung 6.19: Gemessene (schwarze Kästchen) und simulierte (Linien) Frequenzen *f* des Ringoszillators in Abhängigkeit von der Temperatur zwischen Raumtemperatur und 450 °C

Mit steigender Temperatur sinkt die Frequenz des Ringoszillators, da die Ladungsträgerbeweglichkeit mit höherer Temperatur kleiner wird. Zwar sinkt auch der Betrag der Schwellenspannung mit steigender Temperatur, was bei Betrachtung der Simulationen C1 und C4 zu einer höheren Frequenz führen würde, aber nach Gleichung 6.30 und 6.31 wirkt sich die Abnahme der Ladungsträgerbeweglichkeiten stärker aus als die reduzierte Schwellenspannung. Bei Temperaturen unterhalb von 250 °C beträgt die Frequenzänderung ungefähr 24 kHz pro Kelvin, bei Temperaturen oberhalb von 250 °C etwa 11 kHz pro Kelvin. Ein linearer Fit für die Messwerte von 225 °C bis 275 °C liefert eine Frequenzänderung von circa 16 kHz pro Kelvin. Die Frequenz bei 250 °C ist mit 7,3 MHz etwa halb so groß wie die Frequenz bei Raumtemperatur (f = 12,7 MHz), die Frequenz bei 450 °C beträgt mit 5,0 MHz etwas mehr als ein Drittel der Raumtemperaturfrequenz.

Die Messkurve liegt zwischen den beiden Simulationskurven und die Messwerte nehmen von Raumtemperatur bis circa 400 °C ungefähr im gleichen Maße wie die simulierten Werte ab. Ab 400 °C knicken die Simulationskurven aber ein und für Temperaturen oberhalb von 425 °C ist keine Simulation mehr möglich. Der Grund dafür ist die Tatsache, dass bei einer Temperatur zwischen 400 °C und 450 °C die intrinsische Ladungsträgerkonzentration n_i größer wird als die Kanaldotierung $N_{a,d}$. Simulationen für $n_i > N_{a,d}$ sind mit der verwendeten Software nicht möglich. Trotz des intrinsischen Verhaltens oszilliert der Ringoszillator aber weiter. Ausschlaggebend dafür ist die Tatsache, dass auch bei 450 °C der NMOS-Transistor bei $U_1 = U_{dd}$ (siehe Abb. 6.17) leitet und der PMOS-Transistor sperrt und umgekehrt bei U_1 = 0 V der PMOS-Transistor leitet und der NMOS-Transistor sperrt. In Abbildung 6.1 (Abschnitt 6.1.2) ist nämlich auch bei 450 °C der Drain-Source-Strom des NMOS-Transistors bei 5 V größer als der betragsmäßige Strom des PMOS-Transistors bei 0 V und umgekehrt bei -5 V der Drain-Source-Strom des PMOS-Transistors betragsmäßig größer als der Strom des NMOS-Transistors bei 0 V.

Das Verhalten der Stromaufnahme des Ringoszillators ändert sich bei steigender Temperatur anders als die Frequenz. Abbildung 6.20 zeigt die Stromaufnahme des Ringoszillators in Abhängigkeit von der Temperatur. Auch hier entsprechen die schwarzen Kästchen den Messwerten und die beiden Linien den Ergebnissen der Simulationen.

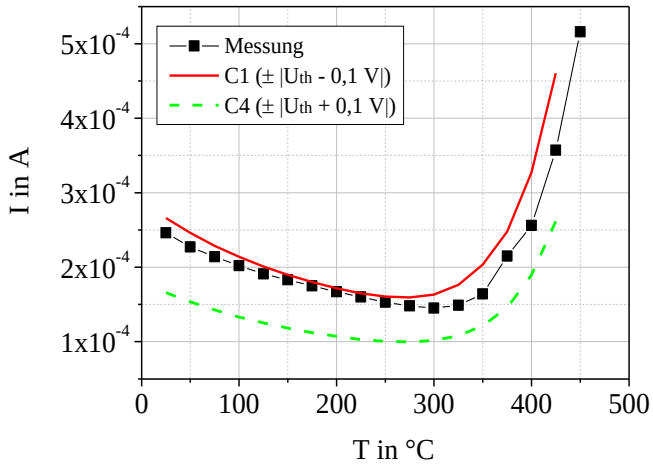


Abbildung 6.20: Gemessene (schwarze Kästchen) und simulierte (Linien) Stromaufnahmen *I* des Ringoszillators in Abhängigkeit von der Temperatur zwischen Raumtemperatur und 450 °C

Bis 300 °C sinkt die Stromaufnahme aufgrund der Reduktion der Ladungsträgermobilität mit steigender Temperatur. Ab einer Temperatur von ungefähr 350 °C steigt der Strom aber stark an. Der Grund dafür sind die Leckströme der 101 CMOS-Inverter, die die Stromaufnahme des Ringoszillators dominieren. Sowohl die NMOS- als auch die PMOS-Transistoren sind für den Leckstrom verantwortlich, da für beide Transistoren das Leckstromniveau in Abhängigkeit von der Temperatur fast gleich ist (vergleiche Abbildung 6.2). In der Simulation ist der Anstieg der Stromaufnahme des Ringoszillators ebenso sichtbar, und die gemessenen Werte liegen wie auch bei den Frequenzen zwischen den Simulationskurven.

Die Ergebnisse aus den Untersuchungen der Ringoszillatoren weisen darauf hin, dass der Leckstrom der Parameter ist, der bei Temperaturen oberhalb von 250 °C Schaltungen am stärksten beeinflusst.

Langzeitstabilität von Frequenz und Stromaufnahme

Neben der Temperaturabhängigkeit von Frequenz und Stromaufnahme wurde das Langzeitverhalten der Ringoszillatoren bei 250 °C und 350 °C beobachtet[16]. Dazu wurden mehrere Ringoszillatoren in Betrieb gelagert und die Änderung von Frequenz und Stromaufnahme in regelmäßigen Abständen überprüft. Die Versorgungsspannung betrug immer U_{dd} = 5 V. Abbildung 6.21 zeigt die Frequenzänderung eines Ringoszillators in Abhängigkeit von der Zeit als absolute Frequenzwerte (linke *y*-Achse) und als prozentuale Änderung bezogen auf den zweiten Messpunkt nach t = 15 h (rechte y- Achse). Die erste Frequenzmessung nach t = 0 h direkt nach dem Schließen des Ofens, aus Darstellungsgründen bei t = 10 h eingezeichnet, zeigt einen deutlich höheren Frequenzwert. Der Unterschied der Frequenz zwischen t = 0 h und t = 15 h beträgt ca. 63 kHz und entspricht einem Temperaturunterschied von ungefähr 4 °C. Er ist darauf zurückzuführen, dass die Temperatur des Ofens bei t = 0 h noch nicht stabil war. Aus diesem Grund ist die prozentuale Frequenzänderung auf den Frequenzwert nach t = 15 h bezogen. Weitere Untersuchungen haben ergeben, dass die Temperatur des Ofens nach etwa zwanzig Minuten stabil ist und sich die Frequenz von t = 0,3 h zu t = 15 h nicht wesentlich ändert.

[16] Eine Langzeituntersuchung bei 450 °C war auch hier aufgrund der Oxidationsproblematik (siehe Kapitel 5) nicht möglich.

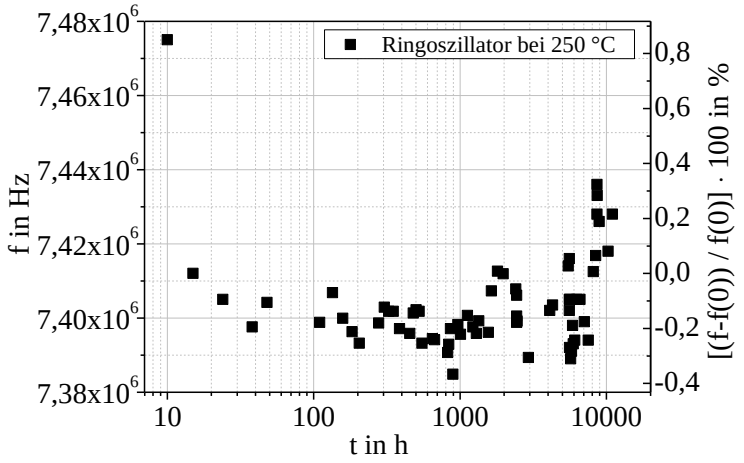


Abbildung 6.21: Frequenz eines Ringsoszillators in Betrieb bei 250 °C in Abhängigkeit von der Zeit als absolute Frequenzwerte (linke y- Achse) und als prozentuale Änderung, bezogen auf den Frequenzwert $f(0)$ nach t = 15 h (rechte y-Achse); t = 0 h ist wegen der logarithmischen Darstellung bei t = 10 h eingezeichnet worden

In den ersten 7500 h schwankt die Frequenz des Ringoszillators um weniger als 40 kHz, d.h. um weniger als 0,4 % der Referenzfrequenz $f(0) = f(15\text{h})$. Diese Schwankung ist auf die Temperaturungenauigkeit des Ofens zurückzuführen. Laut Herstellerangabe schwankt bei einer eingestellten Temperatur von 250 °C die tatsächliche Ofentemperatur an einer festen Position im Ofen um etwa ± 3 K. Dies entspricht mit den Ergebnissen aus Diagramm 6.19 einer Frequenzschwankung von ungefähr ± 48 kHz. Die beobachtete Schwankung bis 7500 h kann damit erklärt werden. Nach 7500 h steigt die Frequenz des Ringoszillators leicht an, bleibt dann aber bei den weiteren Messwerten auf dem erreichten Niveau stabil. Der Grund für den zwischenzeitlichen Frequenzanstieg sind Störungen (Übersprechen) von anderen Bauelementen. Der leichte Anstieg ist im Rahmen der Messgenauigkeit nicht signifikant.

Die Frequenz nach 10.000 h hat sich damit im Vergleich zur Referenzfrequenz $f(0)$ im Rahmen der Messgenauigkeit (± 48 kHz) nicht verändert. Weitere, wegen messtechnischen Problemen in den ersten 1000 h hier nicht dargestellte Ringoszillatoren, zeigten in einem Zeitraum von ca. 2500 h auch keine maßgeblichen Frequenzänderungen[17].

Abbildung 6.22 präsentiert die Entwicklung der Stromaufnahme des Ringoszillators aus Abbildung 6.21 in Abhängigkeit von der Zeit als absolute Stromwerte (linke y-Achse) und als prozentuale Änderung, bezogen auf den zweiten Messpunkt nach t = 15 h (rechte y-Achse). Wie auch bei der Frequenz ergaben weitere Messungen, dass sich die Stromaufnahme von t = 0,3 h zu t = 15 h nicht wesentlich ändert. Aus Darstellungsgründen ist auch hier der Messwert von t = 0 h bei t = 10 h eingezeichnet worden.

In den ersten 7500 h (wieder abgesehen vom ersten Messwert) ist die Stromaufnahme relativ stabil. Nach etwa 7500 h sinkt der Strom kurzzeitig ab und stabilisiert sich dann wieder, weil auch hier zwischenzeitlich andere Bauelementen einen störenden Einfluss haben. Die Stromaufnahme der hier nicht dargestellten Ringoszillatoren zeigt ein entsprechendes Bild.

[17] Nach 2500 h sind diese Ringoszillatoren keineswegs defekt, bisher wurde aber nur eine Lagerungszeit von 2500 h erreicht.

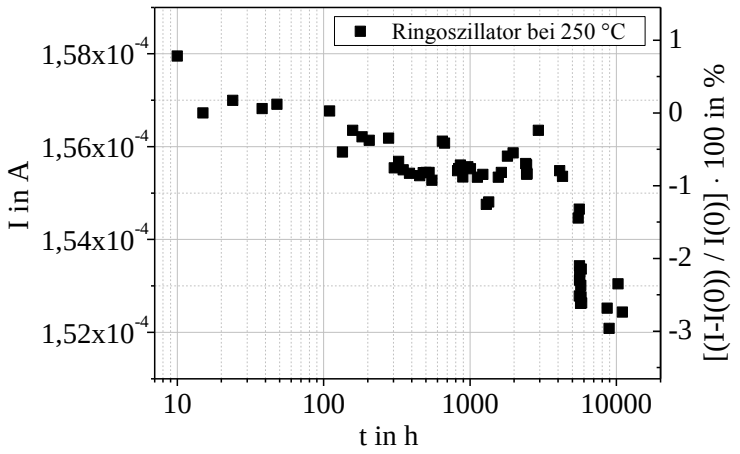


Abbildung 6.22: Stromaufnahme eines Ringsoszillators in Betrieb bei 250 °C in Abhängigkeit von der Zeit als absolute Stromwerte (linke *y*-Achse) und als prozentuale Änderung, bezogen auf den Stromwert *I*(0) nach *t* = 15 h (rechte *y*-Achse); *t* = 0 h ist wegen der logarithmischen Darstellung bei *t* = 10 h eingezeichnet worden

In Ergänzung zu der zeitlichen Entwicklung bei 250 °C ist in Abbildung 6.23 die Veränderung der Frequenz in Abhängigkeit von der Zeit bei einer Ofentemperatur von 350 °C dargestellt. Auf der linken *y*-Achse sind die absoluten Frequenzwerte, auf der rechten y-Achse ist die prozentuale Änderung bezogen auf den zweiten Messpunkt nach *t* = 0,17 h dargestellt. Der erste Messpunkt bei *t* = 0 h (bei *t* = 0,1 h eingezeichnet) hat auch hier wegen der noch andauernden Temperaturstabilisierung des Ofens eine deutlich höhere Frequenz als die weiteren Messpunkte, danach bleibt die Frequenz aber stabil. Innerhalb des Beobachtungszeitraums von circa 1000 h schwankt die Frequenz um weniger als 40 kHz, was auch hier wieder durch die Temperaturungenauigkeit des Ofens erklärt werden kann. Eine Degradation innerhalb dieser 1000 h ist nicht zu erkennen.

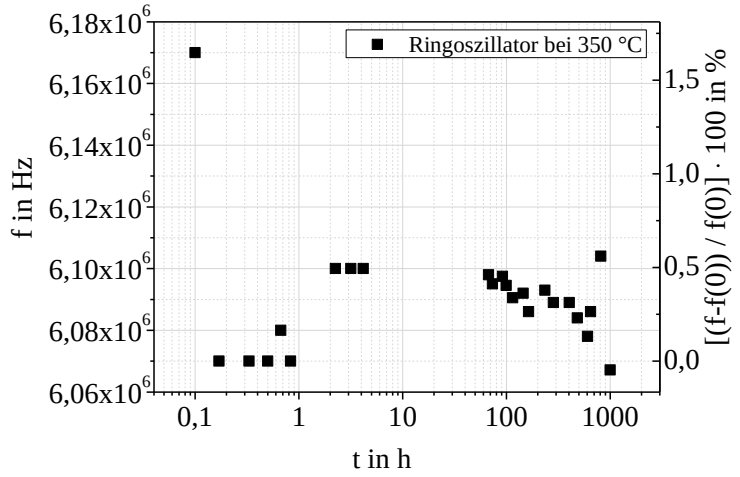


Abbildung 6.23: Frequenz eines Ringsoszillators in Betrieb bei 350 °C in Abhängigkeit von der Zeit als absolute Frequenzwerte (linke *y*-Achse) und als prozentuale Änderung, bezogen auf den Frequenzwert *f*(0) nach *t* = 0,17 h (rechte *y*-Achse); *t* = 0 h ist wegen der logarithmischen Darstellung bei *t* = 0,1 h eingezeichnet worden

Fazit

Aus den durchgeführten Messungen und Simulationen lässt sich schließen, dass die Realisierung digitaler Schaltungen wie die eines Ringoszillators bis 400 °C prinzipiell möglich ist. Auch wenn bei so hohen Temperaturen die Schwellenspannungen und Ladungsträgerbeweglichkeiten für eine um fast einen Faktor drei reduzierte Frequenz gegenüber Raumtemperatur sorgen, ist eine generelle Funktionalität möglich. Ab Temperaturen von 350 °C und mehr sorgen aber zunehmende Leckströme für Einschränkungen. Eine Langzeittätigkeit von Ringoszillatoren konnte bei 250 °C für eine Dauer von 10.000 h und bei 350 °C für einen Zeitraum von 1000 h ohne signifikante Degradation gewährleistet werden.

6.3 Bandgap-Referenzen als analoge Grundschaltungen

6.3.1 Theoretische Aspekte zu Bandgap-Referenzen

Die Bandgap-Referenz-Schaltung (auch Bandgap-Referenz oder Bandgap-Schaltung genannt) ist eine grundlegende analoge Schaltung, deren Ausgangssignal, die Referenzspannung U_{ref}, der Bandlücke des verwendeten Halbleitermaterials entspricht, in diesem Fall Silizium (E_{g} = 1,17 eV bei 0 K). Damit die Referenzspannung über einen definierten Temperaturbereich stabil bleibt, ist eine akkurate Temperatur- bzw. Spannungskompensation notwendig.

Die Funktionalität einer Bandgap-Referenz-Schaltung wird durch die in ihr verwendeten Dioden und Widerstände bestimmt. Die Flussspannung einer Diode U_{Diode} setzt sich aus einem konstanten Anteil $U_{\mathrm{Eg,0}}$, der von der Bandlücke bei 0 K bestimmt wird, einem linear von der Temperatur abhängigen Anteil $U_{\mathrm{D,lin}}$ und einem kleinen nichtlinear temperaturabhängigen Anteil $U_{\mathrm{D,nlin}}$ zusammen.

$$U_{\mathrm{Diode}} = U_{\mathrm{Eg,0}} + U_{\mathrm{D,lin}} + U_{\mathrm{D,nlin}} \qquad [6.33]$$

Der nichtlineare Anteil ist klein und kann in erster Näherung vernachlässigt werden. Es handelt sich dann um eine Bandgap-Referenz erster Ordnung.

Der Zusammenhang zwischen dem Strom I_{D} und der Spannung U_{Diode} der Diode in Durchlassrichtung wird von der Shockley-Gleichung beschrieben:

$$I_{\mathrm{D}} = I_{\mathrm{S}} \cdot \left[\exp\left(\frac{q \cdot U_{\mathrm{Diode}}}{k \cdot T} \right) - 1 \right] \qquad [6.34]$$

Dabei ist I_{S} der Sättigungssperrstrom, q die Ladung eines Elektrons bzw. Lochs, k die Boltzmann-Konstante und T die Temperatur. Für den Fall, dass die Spannung U_{Diode} viel größer ist als (kT)/q, gilt in erster Näherung:

$$I_{\mathrm{D}} \approx I_{\mathrm{S}} \cdot \exp\left(\frac{q \cdot U_{\mathrm{Diode}}}{k \cdot T} \right) \qquad [6.35]$$

Damit gilt für den linearen Abteil der Diodenspannung $U_{\mathrm{D,lin}}$:

$$U_{\mathrm{D,lin}} = \frac{k \cdot T}{q} \cdot \ln\left(\frac{I_{\mathrm{D}}}{I_{\mathrm{S}}} \right) \qquad [6.36]$$

Abbildung 6.24 zeigt ein vereinfachtes Schema der untersuchten Bandgap-Referenz. Die Diode D2 besteht aus 15 parallel geschalteten Dioden der Fläche A_{D1} von Diode D1. Damit ist die Diode D2 mit ihrer Fläche A_{D2} 15-mal so groß wie die Diode D1, hat aber eine 15-mal so kleine Stromdichte. Der lineare Anteil der Flussspannung von Diode D1 hat deshalb einen anderen Temperaturkoeffizienten als die Flussspannung von Diode D2. Die Differenz der beiden Spannungen $\Delta U_{\mathrm{Diode}} = U_{\mathrm{Diode1}} - U_{\mathrm{Diode2}}$ nimmt mit der Temperatur linear zu. Ziel ist es, mit dieser linearen Spannung den negativen linearen Anteil in der Diodenspannung zu kompensieren (vergleiche Abbildung 6.25).

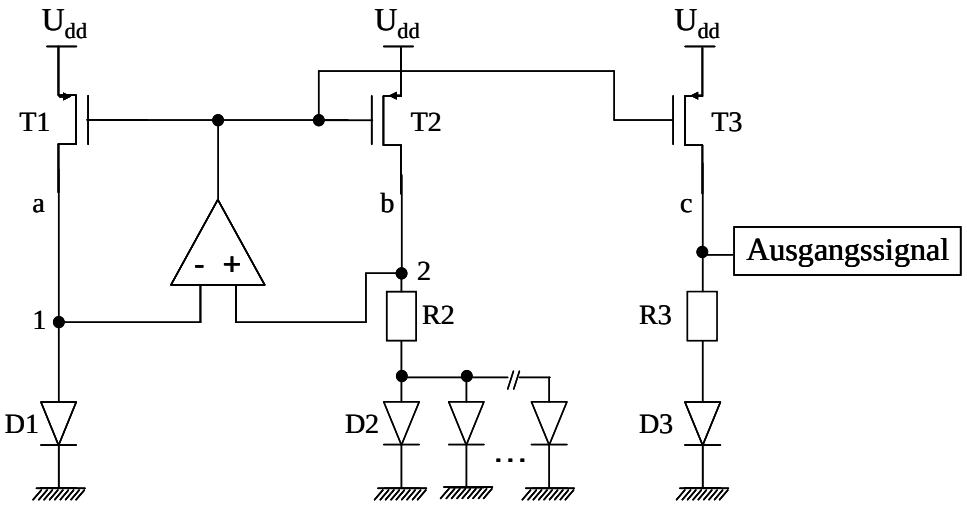


Abbildung 6.24: Schema der untersuchten Bandgap-Referenz mit den Dioden D1, D2 und D3 sowie den Widerständen R2 und R3; U_{dd} ist die Versorgungsspannung

Der Operationsverstärker in Abbildung 6.24 dient dazu, an den Knoten 1 und 2 die gleiche Spannung zu erzeugen. Die Differenzspannung ΔU_{Diode} fällt dann über dem Widerstand R2 ab und in den Pfaden a, b und c fließt, bedingt durch den Stromspiegel T1 bis T3, jeweils der gleiche Strom, nämlich I_{PTAT} ([18]).

$$I_{\text{PTAT}} = \frac{\Delta U_{\text{Diode}}}{R_2} = \frac{\frac{k \cdot T}{q} \cdot \ln\left(\frac{A_{\text{D2}}}{A_{\text{D1}}}\right)}{R_2} \qquad [6.37]$$

Über dem Widerstand R3 fällt dann die Spannung U_{PTAT} ab.

$$U_{\text{PTAT}} = I_{\text{PTAT}} \cdot R_3 \qquad [6.38]$$

Die Diode D3 hat die gleiche Größe wie die Diode D1, d.h. auch an der Diode D3 fällt die Spannung U_{Diode1} ab. Am Ausgang ergibt sich dann insgesamt die Spannung U_{ref} [BAN99]:

$$U_{\text{ref}} = U_{\text{PTAT}} + U_{\text{Diode1}} \qquad [6.39]$$

Da diese Spannung wegen Gleichung 6.37 vom Verhältnis der beiden Widerstände R2 und R3 abhängt, können bei festgelegten Diodenflächen die Widerstände so gewählt werden, dass U_{PTAT} den linearen Anteil U_{lin} von U_{Diode1}, kompensieren kann[19]. Übrig bleibt neben dem kleinen nichtlinearen Anteil, der in Abbildung 6.25 als gestrichelt gezeichneter Bogen zu erkennen ist, die temperaturunabhängige Spannung $U_{\text{Eg,0}}$, die von der Bandlücke bei 0 K bestimmt wird.

[18] Die Abkürzung „PTAT“ steht für den englischen Ausdruck „proportional to absolute temperature“ und bezeichnet die Tatsache, dass dieser Strom linear mit der Temperatur ansteigt.

[19] Es ist zu beachten, dass $U_{\text{D,lin}}$ wegen $I_{\text{D}} < I_{\text{S}}$ einen negativen Temperaturkoeffizienten hat und U_{PTAT} wegen $A_{\text{D2}} > A_{\text{D1}}$ einen positiven Temperaturkoeffizienten.

Die Temperaturabhängigkeit der Bandlücke kann nach [MUL86] und [WER96] vernachlässigt werden (vergleiche auch Abschnitt 6.1.1), so dass im Temperaturbereich bis 450 °C bei korrekter Spannungskompensation eine konstante Referenzspannung zu erwarten wäre.

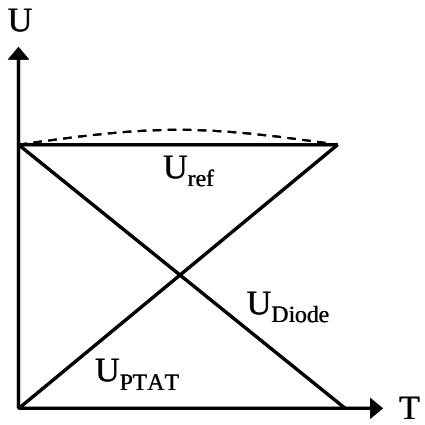


Abbildung 6.25: Schematische Darstellung der Spannungskompensation der Flussspannung U_{Diode} durch die Spannung U_{PTAT} zur Referenzspannung U_{ref}; der gestrichelt gezeichnete Bogen kennzeichnet den kleinen nichtlinearen Anteil an der Dioden-Flussspannung

Die Temperaturstabilität der Bandgap-Referenz hängt also wesentlich von der Funktionalität des Operationsverstärkers und der verwendeten Dioden, Transistoren und Widerstände ab.

6.3.2 Charakterisierung und Langzeitverhalten von Bandgap-Referenzen

Um die in Abschnitt 6.4.1 präsentierte Bandgap-Schaltung zu charakterisieren, wurden ihre Referenzspannung U_{ref} und die Stromaufnahme I bei Temperaturen zwischen Raumtemperatur und 450 °C gemessen. In Abbildung 6.26 sind die Ergebnisse für die Spannungsabhängigkeit von der Temperatur dargestellt (siehe auch [GRE12]).

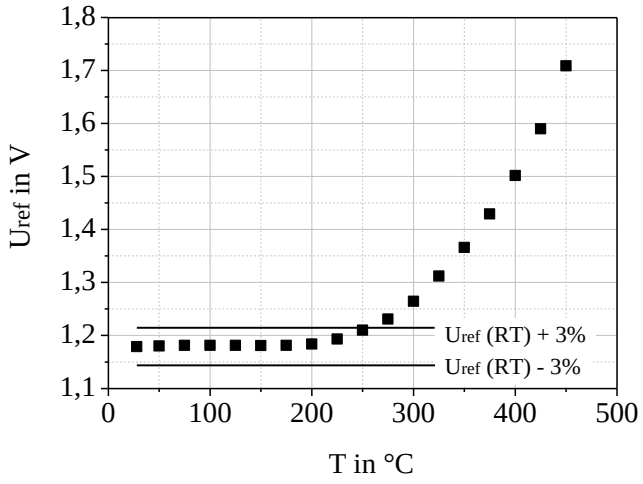


Abbildung 6.26: Referenzspannung U_{ref} der Bandgap-Schaltung in Abhängigkeit von der Temperatur zwischen Raumtemperatur und 450 °C; die beiden schwarzen Linien markieren den akzeptierten Toleranzbereich der Spannungsschwankung von ± 3 %; U_{dd} = 4 V

Bei den Messungen war die Versorgungsspannung der Schaltung U_{dd} = 4 V. In Vergleichsmessungen bei mehreren Versorgungsspannungen zwischen 4 V und 6 V zeigte sich, dass sich die Referenzspannung und die Stromaufnahme bei steigender Temperatur umso mehr erhöhen, je höher die Versorgungsspannung ist.

Die Bandgap-Referenzspannung beträgt bei Raumtemperatur etwa 1,18 V und circa 1,21 V bei 250 °C. Dies entspricht einer Änderung von weniger als 3 %. Für Temperaturen oberhalb von 250 °C steigt die Spannung aber stark an und erreicht mehr als 1,7 V bei 450 °C. Die Temperaturkompensation funktioniert für Temperaturen größer als 250 °C nicht mehr. Der Grund dafür liegt in der Schwellenspannung der verwendeten PMOS-Transistoren. Die Schwellenspannung der PMOS-Transistoren der Bandgap-Referenz liegt im Unterschied zu den im Ringoszillator verwendeten PMOS-Transistoren bei Raumtemperatur nicht bei -2 V sondern bei etwa -1 V und steigt für Temperaturen über 250 °C über 0 V. Der Operationsverstärker ist dann nicht mehr in der Lage, die Spannung zwischen den Knoten 1 und 2 auszugleichen und der Strom in den Pfaden a und b ist nicht mehr gleich. Somit verändert sich auch die Spannung U_{PTAT}, die den linearen Anteil der Dioden-Flussspannung nun nicht mehr korrekt kompensieren kann. Es muss aber angemerkt werden, dass beim Entwurf der Schaltung eine Funktionalität oberhalb von 250 °C nicht geplant war.

Eine Möglichkeit, die Referenzspannung auch für Temperaturen oberhalb von 250 °C stabil zu halten, wäre die Verwendung der PMOS-Transistoren mit einem Schwellenspannungswert von -2 V bei Raumtemperatur. Diese hätten sogar bei 450 °C noch eine negative Schwellenspannung, wie aus den Ergebnissen von Abbildung 6.4 folgt.

In Abbildung 6.27 ist die Änderung der Stromaufnahme der Bandgap-Referenz mit der Temperatur aufgetragen. Die Kurve kann in zwei Abschnitte eingeteilt werden. Bis zu einer Temperatur von ungefähr 250 °C steigt der Strom linear an, nimmt dann aber für höhere Temperaturen stärker zu. Auch dieser zweite Anstieg ist linear, aber mit einer um ungefähr einen Faktor vier höheren Steigung als bis 250 °C. Der Grund für den erhöhten Stromfluss ist vermutlich auch hier ein erhöhter Leckstrom. Ob dieser vom Transistor im Pfad c, von der Diode D3 oder von beiden Bauelementen kommt, kann mit der Bandgap-Schaltung allein nicht festgestellt werden.

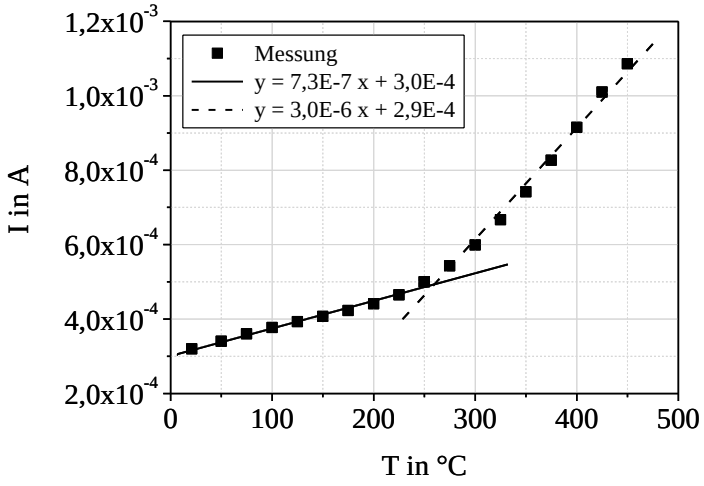


Abbildung 6.27: Stromaufnahme *I* der Bandgap-Schaltung in Abhängigkeit von der Temperatur zwischen Raumtemperatur und 450 °C; die Stromänderung knickt bei ungefähr 250 °C nach oben ab; im Bereich zwischen 250 °C und 450 °C (gestrichelte Linie) ist die Steigung etwa viermal so groß wie die Steigung zwischen Raumtemperatur und 250 °C (durchgezogene Linie); U_{dd} = 4 V

In Ergänzung zu der Untersuchung des Temperaturverhaltens wurde die Langzeitstabilität der Bandgap-Schaltung bei 250 °C betrachtet. Da die Referenzspannung bei 350 °C schon weit außerhalb des Toleranzbereiches von ± 3 % liegt, wurde keine Langzeitanalyse bei 350 °C durchgeführt. Abbildung 6.28 zeigt die Ergebnisse der Untersuchung. Der Mittelwert der Referenzspannung schwankt im Beobachtungszeitraum um weniger als 0,5 % und kann damit als stabil angesehen werden. Allerdings ist zu bemerken, dass die einzelnen Bandgap-Referenzen mit bis zu 80 mV Differenz sehr weit auseinander liegen, was eine Folge von Mismatch und der Prozessabhängigkeit ist.

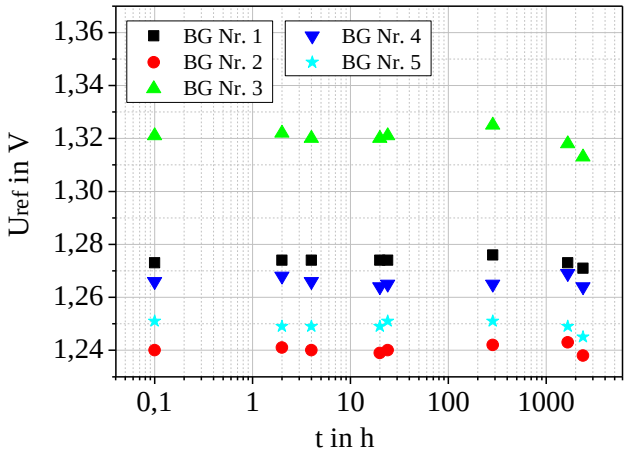


Abbildung 6.28: Referenzspannungen von fünf Bandgap-Schaltungen (Abkürzung: BG) in Betrieb bei 250 °C in Abhängigkeit von der Zeit; t = 0 h ist wegen der logarithmischen Darstellung bei t = 0,1 h eingezeichnet; alle Bandgap-Schaltungen waren vorher schon circa 3000 h bei 250 °C in Betrieb

Fazit

Insgesamt lässt sich aus den Ergebnissen der Untersuchungen der Bandgap-Referenzen schließen, dass eine analoge Schaltungsfunktion bis 250 °C im Rahmen definierter Abweichungen möglich ist. Über einen Zeitraum von mehreren tausend Stunden bei 250 °C konnte bei den untersuchten Bandgap-Referenzen im Rahmen der Messgenauigkeiten keine Degradation der Referenzspannung festgestellt werden. Eine Erweiterung des Temperaturbereichs bis 300 °C oder sogar bis 350 °C könnte durch Anpassungen wie der Wahl geeigneter PMOS-Transistoren erreicht werden.

Kapitel 7

Zusammenfassung und Ausblick

Die vorangegangenen Kapitel geben einen umfassenden Überblick über die wichtigsten Zuverlässigkeitsphänomene und Ausfallmechanismen, die CMOS-Bauelemente und CMOS-Schaltungen auf SOI bei Temperaturen von 250 °C und mehr betreffen können. Dabei wurde die Vorgehensweise der Analysen und Auswertungen so weit wie möglich an die JEDEC-Standards angelehnt. In diesem Kapitel sollen nun die Ergebnisse noch einmal zusammengefasst und ein Ausblick für weitere Untersuchungen bezüglich der Zuverlässigkeit von Hochtemperaturprozessen gegeben werden.

Gateoxid

Zur Untersuchung der Qualität und Zuverlässigkeit des Gateoxids wurden TDDB-Messungen an Gateoxidkondensatoren durchgeführt. Dabei wurde eine Abhängigkeit von der Polarisation des elektrischen Feldes festgestellt. Wenn bei der Messung Elektronen von der Gateelektrode her in das Oxid injiziert werden, liegen kurze Durchbruchzeiten und extrinsische Ausfälle vor und in IU-Kennlinien treten erhöhte Stromwerte auf („Buckel“). Der Grund dafür liegt in der rauen Struktur des Polysiliziums, die für Defekte im Oxid an der Polysilizium-Siliziumdioxid-Grenzfläche verantwortlich ist.

Für TDDB-Messungen an PMOS-Kondensatoren in Akkumulation stimmen die ermittelten Feldbeschleunigungsfaktoren und Aktivierungsenergien mit Angaben in der Literatur überein. Es wurde gezeigt, dass die bis 200 °C bekannten Gesetzmäßigkeiten unverändert auch für höhere Temperaturen bis 350 °C gelten. Es treten bei diesen hohen Temperaturen also keine zusätzlichen Ausfallmechanismen auf.

EEPROM-Speicherzellen

Bezüglich der Untersuchung des Datenerhalts an EEPROM-Speicherzellen konnte festgestellt werden, dass bei Lagerungstemperaturen oberhalb von 250 °C zusätzliche Ausfallmechanismen auftreten, die bei niedrigeren Testtemperaturen in den üblichen Lagerungszeiten von einigen Tausend Stunden nicht erkennbar sind. Vor allem gelöschte Zellen sind von diesem Effekt betroffen. Der Grund für den zusätzlichen Ladungsverlust sind positive mobile Ionen, die ihren Ursprung in der Prozessierung des Back-Ends haben. Die Aktivierungsenergie für diesen Effekt beträgt etwa 2 eV.

Wichtig ist anzumerken, dass sowohl die Festlegung des Ausfallkriteriums (Größe der prozentualen Änderung) als auch der Temperaturbereich, in dem die Untersuchungen vorgenommen werden, eine entscheidende Rolle bei der Berechnung der Aktivierungsenergie und der Lebensdauer spielen. Für schnelle Vergleiche von EEPROM-Zellen auf verschiede-

nen Wafern ist zudem eine Lagerungstemperatur von 350 °C gut geeignet, weil schon innerhalb von 24 h Unterschiede deutlich werden.

Die Lagerung der EEPROM-Zellen beeinflusst die auf dem Floating-Gate gespeicherte Ladung, zerstört die Zellen an sich aber nicht. Auch eine mehrere Stunden bei 450 °C gelagerte Zelle kann weiterhin programmiert und gelöscht werden. Eine erneute Lagerung wird aber von den vorherigen Lagerungsexperimenten beeinflusst. Deshalb ist es wichtig, dass Messungen zum Datenerhalt nur an Wafern vorgenommen werden, die vorher noch keine Temperaturlagerung erfahren haben.

Bei den Untersuchungen zur Zyklenfestigkeit der EEPROM-Zellen wurde festgestellt, dass der Verlauf des Programmierfensters mit zunehmender Zyklenzahl von der Temperatur, der Programmierspannung und der Tunneloxiddicke abhängt. Je höher die Temperatur ist, bei der eine Zelle gezykelt wird, desto schneller ist die Zelle defekt. Bei der in einer Schaltung üblichen Programmierspannung von U_{pp} = 16 V konnten die untersuchten EEPROM-Zellen bei 25 °C mindestens 1 Million Mal, bei 250 °C mehr als 10.000-mal und bei 400 °C immer noch mehr als 1000-mal gezykelt werden, bevor die gelöschte und die programmierte Schwelle nicht mehr zu unterscheiden waren.

Metallisierung

Elektromigration spielt bei der Wolframmetallisierung des H10-Prozesses keine Rolle, so wie es auch in der Literatur beschrieben wird. Sind in Messungen Widerstandszunahmen zu beobachten, sind diese immer auf die Degradation des Pad-Messnadel-Kontaktes durch die Oxidation der Pads zurückzuführen.

In den Stressmigrationstests konnten während 10.000 h bei 250 °C und einigen Hundert Stunden bei 350 °C an Metallbahnen, Metallmäandern und Viaketten keine Widerstandsänderungen festgestellt werden. Änderungen von Kontakt- und Siliziumwiderständen nach einer Temperaturlagerung sind eine Folge aufgebrochener Wasserstoffbindungen an den Korngrenzen. Vor allem Poly-Widerstände und Poly-Metall-Kontakte sind davon betroffen. Die Größenordnung der Änderungen ist aber insgesamt unkritisch.

Transistoren und Testschaltungen

Der Betrag der Schwellenspannung von NMOS- und PMOS-Transistoren nimmt bis etwa 275 °C annähernd linear mit der Temperatur ab. Für größere Temperaturen bis 400 °C ist die Abnahme auch linear, die Änderung pro Kelvin ist aber etwa dreimal so groß. Der Übergang zwischen den beiden Abschnitten entspricht dem Übergang von einem Transistor mit einem vollständig verarmten Siliziumfilm zu einem nur teilweise verarmten Film. Während der Betrag des Sättigungsstromes mit der Temperatur abnimmt, steigt der Betrag des Leckstromes mit der Temperatur an. Der Übergang von vollständiger Verarmung zu unvollständiger Verarmung ist auch hier zu erkennen. Prinzipiell sind die Transistoren bis etwa 400 °C funktionsfähig.

Die beiden hauptsächlichen Zuverlässigkeitsphänomene, die Transistoren im Betrieb betreffen können, sind der Einfluss „heißer“ Ladungsträger und der Einfluss einer negativen Gate-Source-Spannung. Hot Carrier - Effekte spielen bei 250 °C keine Rolle. Bei Raumtemperatur ist bei PMOS-Transistoren auf den Einfluss des Back-Gates zu achten, denn bei

einer großen Potentialdifferenz zwischen Drain und Back-Gate erfolgt eine Hot Carrier - Schädigung nicht nur an der Grenzfläche zwischen Film und Gateoxid (Si-SiO_2), sondern auch an der Grenzfläche zwischen Film und vergrabenem Oxid (Si-BOX). Dieser Aspekt ist für SOI-Transistoren sehr wichtig, zumal er in den für Bulk-Transistoren ausgerichteten JEDEC-Standards nicht explizit erwähnt wird.

Bei den Messungen zur Parameterinstabilität bei negativer Gate-Source-Spannung konnten zwei wichtige Erkenntnisse gewonnen werden. Es ist möglich, statt der Änderung der Schwellenspannung die Änderung des Drain-Source-Stromes zu betrachten, denn beide Änderungen folgen den gleichen Gesetzmäßigkeiten bezüglich ihrer Abhängigkeit von der Zeit, der Stressspannung, der Temperatur und der Ausheilung. Zudem gelten die Gesetzmäßigkeiten nicht nur bis 200 °C, sondern auch bis 300 °C. Bei Betrachtung der Stromänderung kann das sonst übliche Problem der zwischenzeitlichen Ausheilung umgangen werden, so dass diese Methode eine sinnvolle Alternative zu der im JEDEC-Standard beschriebenen Vorgehensweise darstellt.

Weiterhin konnte festgestellt werden, dass Ringoszillatoren bis 400 °C funktionstüchtig sind, die Stromaufnahme aber spätestens ab 300 °C durch hohe Leckströme beeinträchtigt wird. Bei 250 °C ändern sich Frequenz und Stromaufnahme innerhalb von 10.000 h im Rahmen der Messgenauigkeiten nicht. Die Bandgap-Referenzen hingegen sind nur bis etwa 250 °C nutzbar, da hier für höhere Temperaturen Abweichungen der Referenzspannung nicht mehr vermeidbar sind. Mit entsprechenden Schaltungskonzepten besteht aber die Annahme, dass auch analoge Schaltungen auf SOI bis zu 300 °C oder sogar 350 °C realisierbar sind.

Ausblick

Insgesamt gibt es zwei Aspekte, die unter Umständen kritisch sind. Dies sind die extrinsischen Ausfälle und frühen Durchbruchspannungen der Gateoxidkondensatoren und der stark von der Back-End-Prozessierung abhängende Datenerhalt der EEPROM-Zellen. Für beide Phänomene sind weitere Analysen notwendig, um die damit verbundenen Zuverlässigkeitsprobleme einzuschränken.

Die vorgestellten Zuverlässigkeitsuntersuchungen und Charakterisierungen wurden an einem CMOS-Prozess auf SOI mit einer minimalen Transistor-Kanallänge von 1 µm durchgeführt, so dass die gewonnenen Erkenntnisse eigentlich auch nur für Bauelemente, die in diesem Prozess gefertigt wurden, gültig sind. Eine Möglichkeit, die Anwendung der Untersuchungsprozeduren auch an einem anderen CMOS-Prozess auf SOI zu testen und damit in Richtung einer größeren Allgemeingültigkeit zu führen, sind weitere Untersuchungen am zweiten Hochtemperatur-SOI-CMOS-Prozess des Fraunhofer IMS (siehe auch Kapitel 2).

In diesem „H035“ genannten Prozess gibt es einige Änderungen in Bezug auf den H10-Prozess, durch die auch die Zuverlässigkeitsaspekte eine andere Relevanz erhalten könnten. Der wichtigste Aspekt ist die Reduzierung der minimalen Strukturgröße von 1 µm auf 0,35 µm, wodurch Leckströme eine noch größere Bedeutung erhalten. Für Hot Carrier - Messungen ergeben sich möglicherweise Vorteile durch eine „Lightly Doped Drain (LDD)“ - Zone, die den Einfluss „heißer“ Ladungsträger reduzieren kann. Mit der Strukturgrößenverringerung der Transistoren gehen aber auch eine Reduzierung der Kontakt- und Viagrößen von 0,8 µm auf 0,4 µm und der minimalen Bahnbreite von 1 µm auf 0,48 µm einher. Stressmigrationsphänomene könnten deshalb kritisch sein. Zudem wird auf die Polysilizium- und Aktivgebietsflächen eine Titan-Silizid-Schicht zur Widerstandsreduzierung aufgebracht.

Auch diese könnte einen Einfluss haben, z. B. auf die Temperaturabhängigkeit der Kontaktwiderstände. Ein weiterer Punkt ist das dünne Gateoxid (10 nm), bei dem Zuverlässigkeitseigenschaften besonders kritisch sind. Für die EEPROM-Zellen muss beachtet werden, dass das Tunneloxid nun nicht mehr durch einen RTA-Schritt sondern in einer langsameren Prozedur in einem Oxidationsofen gefertigt wird. Auch hier ist nicht auszuschließen, dass dies einen (positiven oder negativen) Einfluss auf die Zuverlässigkeit der Speicherzellen hat.

Literaturverzeichnis

[ABE00] R. B. Abernethy, „The New Weibull Handbook“, 4th Edition, (2000)

[ACO96] A. Acovic, G. La Rosa, Y.-C. Sun, „A review of hot-carrier degradation mechanisms in MOSFETs“, Microelectronics Reliability, 36 (7/8), S. 845-869, (1996)

[AHN87] K. Y. Ahn, „A comparison of tungsten film deposition techniques for very large scale integration technology“, Thin Solid Films, 153 (1-3), S. 469-478, (1987)

[AME70] I. Ames, F. M. d'Heurle, R. E. Horstmann, „Reduction of electromigration in aluminum films by copper doping“, IBM Journal of Research and Development, 14 (4), S. 461-463, (1970)

[ARR89] S. Arrhenius, „Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren“, Zeitschrift für physikalische Chemie, 4, S. 226-248, (1889)

[BAG84] D. A. Baglee, „Characteristics and reliability of 100 Å oxides“, Proceedings of the Reliability Physics Symposium, S. 152-155, (1984)

[BAN99] H. Banba, H. Shiga, A. Umezawa, T. Miyaba, T. Tanzawa, A. Atsumi, K. Sakui, „A CMOS bandgap reference circuit with sub-1-V operation“, IEEE Journal of Solid-State Circuits, 34 (5), S. 670-674, (1999)

[BEN53] A. Benard, E. C. Bos-Levenbach, „Het uitzetten van waarnemingen op waarschijnlijkdeids-papier“ (The plotting of observations on probability paper), Statistica Neerlandica, 7, S. 163-173, (1953); Übersetzung von R. Schop, S. 1-8, (2001); http://www.barringer1.com/wa_files/The-plotting-of-observations-on-probability-paper.pdf

[BER06] J. B. Bernstein, M. Gurfinkel, X. Li, J. Walters, Y. Shapira, M. Talmor, „Electronic circuit reliability modeling“, Microelectronics Reliability, 46 (12), S. 1957-1979, (2006)

[BER81] A. Berman, „Time-zero dielectric reliability test by ramp method“, Proceedings of the Reliability Physics Symposium, S. 204-209, (1981)

[BLA67] J. R. Black, „Mass transport of aluminum by momentum exchange with conducting electrons“, Proceedings of the Reliability Physics Symposium, S. 148-159, (1967)

[BLA69] J. R. Black, „Electromigration - A brief survey and some recent results“, IEEE Transactions on Electron Devices, 16 (4), S. 338-347, (1969)

[BLE76] I. A. Blech, „Electromigration in thin aluminum films on titanium nitride“, Journal of Applied Physics, 47 (4), S. 1203-1208, (1976)

[BOY89] K. C. Boyko, D. L. Gerlach, „Time Dependent Dielectric Breakdown of 210Å Oxides“, Proceedings of the International Reliability Physics Symposium, S. 1-8, (1989)

[CAR98] H. Caruso, A. Dasgupta, „A fundamental overview of accelerated-testing analytic models“, Proceedings of the Annual Reliability and Maintainability Symposium, S. 389-393, (1998)

[CHE85] I. C. Chen, S. Holland, C. Hu, „A quantitative physical model for time-dependent breakdown in SiO_2“, Proceedings of the Reliability Physics Symposium, S. 24-31, (1985)

[CHE95] J. Chen, J.-P. Colinge, „Tungsten metallization technology for high temperature silicon-on-insulator devices“, Materials Science and Engineering B, 29 (1-3), S. 18-20, (1995)

[COL97] J.-P. Colinge, „Silicon-on-Insulator Technology: Materials to VLSI“, 2^{nd} Edition, Kluwer Academic Publishers, (1997)

[CRI90] G. Crisenza, G. Ghidini, S. Manzini, A. Modelli, M. Tosi, „Charge loss in EPROM due to ion generation and transport in interlevel dielectric“, IEEE Electron Device Meeting IEDM, S. 107-110, (1990)

[CRI97] S. Cristoloveanu, „Hot-carrier degradation mechanisms in silicon-on-insulator MOSFETs“, Microelectronics Reliability, 37 (7), S. 1003-1013, (1997)

[CRO79] D. L. Crook, „Method of determining reliability screens for time dependent dielectric breakdown“, Proceedings of the Reliability Physics Symposium, S. 1-7, (1979)

[CUR84] J. Curry, G. Fitzgibbon, Y. Guan, R. Muollo, G. Nelson, A. Thomas, „New failure mechanisms in sputtered aluminum-silicon films“, Proceedings of the Reliability Physics Symposium, S. 6-8, (1984)

[DEA80] B. E. Deal, „Standardized terminology for oxide charges associated with thermally oxidized silicon“, IEEE Transactions on Electron Devices, 27 (3), S. 606-608, (1980)

[DEN04] M. Denais, A. Bravaix, V. Huard, C. Parthasarathy, G. Ribes, F. Perrier, Y. Rey-Tauriac, N. Revil, „On-the-fly characterization of NBTI in ultra-thin gate oxide PMOSFET’s“, IEEE Electron Device Meeting IEDM, S. 109-112, (2004)

[DeS99] B. De Salvo, G. Ghibaudo, G. Pananakakis, B. Guillaumot, P. Candelier, G. Reimbold, „A new extrapolation law for data-retention time-to-failure of nonvolatile memories“, IEEE Electron Device Letters, 20 (5), S. 197-199, (1999)

[DeS99b] B. De Salvo, G. Ghibaudo, G. Pananakakis, G. Reimbold, F. Mondond, B. Guillaumot, P. Candelier, „Experimental and theoretical investigation of nonvolatile memory data-retention“, IEEE Transactions on Electron Devices, 46 (7), S. 1518-1524, (1999)

[DiM99] D. J. DiMaria, J. H. Stathis, „Non-Arrhenius temperature dependence of reliability in ultrathin silicon dioxide films“, Applied Physics Letters, 74 (12), S. 1752-1754, (1999)

[DMH10] „Design Manual H10 - 1.0µm SOI CMOS Process for High Temperature Applications“, Version 0.4, (August 2012)

[DUM01] D. J. Dumin, „Oxide wearout, breakdown, and reliability“, International Journal of High Speed Electronics and Systems, 11 (3), S. 617-718, (2001)

[EUZ81] B. Euzent, N. Boruta, J. Lee, C. Jenq, „Reliability aspects of a floating gate E2PROM“, Proceedings of the Reliability Physics Symposium, S. 11-16, (1981)

[FIS02] A. H. Fischer, A. von Glasow, „Electromigration and stressmigration failure mechanism studies in copper interconnects“, SEMI Technology Symposium, Semicon, (2002)

[FLA95] D. Flandre, „Silicon-on-insulator technology for high temperature metal oxide semiconductor devices and circuits“, Materials Science and Engineering B, 29 (1-3), S. 7-12, (1995)

[FLE92] S. Fleischer, P. T. Lai, Y. C. Cheng, „Simplified closed-form trap-assisted tunneling model applied to nitrided oxide dielectric capacitors“, Journal of Applied Physics, 72 (12), S. 5711-5715, (1992)

[FLE93] D. M. Fleetwood, P. S. Winokur, R. A. Reber, T. L. Meisenheimer, J. R. Schwank, M. R. Shaneyfelt, L. C. Riewe, „Effects of oxide traps, interface traps, and „border traps“ on metal-oxide-semiconductor devices“, Journal of Applied Physics, 73 (10), S. 5058-5074, (1993)

[FOW28] R. H. Fowler, L. Nordheim, „Electron emission in intense electric fields“, Proceedings of the Royal Society London A, 119, S. 173-181, (1928)

[FRA92] P. Francis, A. Terao, B. Gentinne, D. Flandre, J.-P. Colinge, „SOI technology for high-temperature applications“, IEEE Electron Device Meeting IEDM, S. 353-356, (1992)

[FRE38] J. Frenkel, „On pre-breakdown phenomena in insulators and electronic semi-conductors“, Physical Review, 54 (8), S. 647–648, (1938)

[GAL01] M. Gall, C. Capasso, D. Jawarani, R. Hernandez, H. Kawasaki, P. S. Ho, „Statistical analysis of early failures in electromigration“, Journal of Applied Physics, 90 (2), S. 732-740, (2001)

[GAS00] P. Gassot, A. Iline, E. De Backer, M. Tack, D. Wellekens, J. Van Houdt, L. Haspeslagh, „Water-assisted positive ion contamination resulting in charge loss in nonvolatile memories“, Proceedings of the 30th European Solid-State Device Research Conference, S. 268-271, (2000)

[GEN97] B. Gentinne, J.-P. Eggermont, D. Flandre, J.-P. Colinge, „Fully depleted SOI-CMOS technology for high temperature IC applications“, Materials Science and Engineering B, 46 (1-3), S. 1-7, (1997)

[GIR65] G. Giralt, B. Andre, J. Simonne, D. Esteve, „Influence de la température sur les dispositifs semiconducteurs du type M.O.S. (Thermal drift of M.O.S. devices)“, Electronics Letters, 1 (7), S. 185-186, (1965)

[GLA05] A. von Glasow, Dissertation zum Thema „Zuverlässigkeitsaspekte von Kupfermetallisierungen in integrierten Schaltungen“, Technische Universität München, (2005)

[GLE97] R. J. Gleixner, B. M. Clemens, W. D. Nix, „Void nucleation in passivated interconnect lines: Effects of site geometries, interfaces, and interface flaws“, Journal of Materials Research, 12 (8), S. 2081-2090, (1997)

[GOG97] D. Gogl, Dissertation zum Thema „Untersuchungen zur Realisierung hochtemperaturtauglicher EEPROM-Speicher in SIMOX-Technologie“, Universität Duisburg, (1997)

[GRA09] T. Grasser, B. Kaczer, W. Goes, T. Aichinger, P, Hehenberger, M. Nelhiebel, „A two-stage model for negative bias temperature instability“, Proceedings of the IEEE International Reliability Physics Symposium, S. 33-44, (2009)

[GRA11] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, M. Toledano Luque, M. Nelhiebel, „The paradigm shift in understanding the bias temperature instability: From reaction-diffusion to switching oxide traps“, IEEE Transactions on Electron Devices, 58 (11), S. 3652-3666, (2011)

[GRA11b] T. Grasser, T. Aichinger, G. Pobegen, H. Reisinger, P.-J. Wagner, J. Franco, M. Nelhiebel, B. Kaczer, „The ‘permanent’ component of NBTI: Composition and annealing“, Proceedings of the IEEE International Reliability Physics Symposium, S. 605-613, (2011)

[GRA99] K.-T. Grasser, Dissertation zum Thema „Mixed-Mode Device Simulation“, Technische Universität Wien, (1999)

[GRE11] K. Grella, H. Vogt, U. Paschen, „High temperature reliability investigations of EEPROM memory cells realised in silicon-on-insulator (SOI) technology“, HiTEN 2011, (2011)

[GRE12] K. Grella, S. Dreiner, A. Schmidt, W. Heiermann, H. Kappert, H. Vogt, U. Paschen, „High temperature characterization up to 450 °C of MOSFETs and basic circuits realized in a silicon-on-insulator (SOI) CMOS-technology“, HiTEC 2012, (2012)

[GRO90] G. Groeseneken, J.-P. Colinge, H. E. Maes, J. C. Alderman, S. Holt, „Temperature dependence of threshold voltage in thin-film SOI MOSFETs“, IEEE Electron Device Letters, 11 (8), S. 329-331, (1990)

[HAI95] M. Hain, H. Körner, B. Neureither, S. Röhl, „A highly reliable, low cost 0.5 µm three level tungsten metallization“, Applied Surface Science, 91 (1-4), S. 374-377, (1995)

[HAR77] E. Harari, „Conduction and trapping of electrons in highly stressed ultrathin films of thermal SiO_2“, Applied Physics Letters, 30 (11), S. 601-603, (1977)

[HCP11] „Handbook of Chemistry and Physics“, 92nd Edition, (2011-2012); http://www.hbcpnetbase.com

[HEH10] P. Hehenberger, H. Reisinger, T. Grasser, „Recovery of negative and positive bias temperature stress in pMOSFETs“, Final report of the IEEE International Integrated Reliability Workshop, S. 8-11, (2010)

[HER11] R. W. Herfst, J. Schmitz, A. J. Scholten, „Simultaneous extraction of threshold voltage and mobility degradation from on-the-fly NBTI measurements“, Proceedings of the IEEE International Reliability Physics Symposium, S. 929–932, (2011)

[HMP99] Hybrid Memory Products Ltd., „Floating Gate Memory Arrays“, S. 1-6, (1999); http://www.hmpeurope.com/pub/app_notes/ANH012A.pdf

[HON95] S. Q. Hong, T. Wetteroth, H. Shin, S. R. Wilson, W. M. Huang, J. Foerstner, M. Racanelli, H. C. Shin, B.-Y. Hwang, D. K. Schroder, „Integrity of gate oxide on TFSOI materials“, Proceedings of the IEEE International SOI Conference, S. 22-23, (1995)

[HON97] S. Q. Hong, T. Wetteroth, H. Shin, S. R. Wilson, D. Werho, T.-C. Lee, D. K. Schroder, „Improvement in gate oxide integrity in thin-film silicon-on-insulator substrates by lateral gettering“, Applied Physics Letters, 71 (23), S. 3397-3399, (1997)

[HOO01] T. B. Hook, E. Adler, F. Guarin, J. Lukaitis, N. Rovedo, K. Schruefer, „The effects of fluorine on parametrics and reliability in a 0.18-µm 3.5/6.8 nm dual gate oxide CMOS technology“, IEEE Transactions on Electron Devices, 48 (7), S. 1346-1353, (2001)

[HOU08] J. van Houdt, R. Degraeve, G. Groeseneken, H. E. Maes, „Physics of Flash Memories“, Kapitel 4.2.2, in „Nonvolatile Memories with Emphasis on Flash“, S. 131-140, (2008)

[HU85] C. Hu, S. C. Tam, F. C. Hsu, P. K. Ko, T. Y. Chan, K. W. Terrill, „Hot-electron-induced MOSFET degradation - model, monitor, and improvement“, IEEE Journal of Solid-State Circuits, 20 (1), S. 295-305, (1985)

[HU99] C. Hu, Q. Lu, „A unified gate oxide reliability model“, Proceedings of the IEEE International Reliability Physics Symposium, S. 47-51, (1999)

[HUA06] V. Huard, M. Denais, C. Parthasarathy, „NBTI degradation: From physical mechanisms to modelling“, Microelectronics Reliability, 46 (1), S. 1-23, (2006)

[HUA10] V. Huard, „Two independent components modeling for Negative Bias Temperature Instability“, Proceedings of the IEEE International Reliability Physics Symposium, S. 33-42, (2010)

[IOA06] D. P. Ioannou, R. Mishra, D. E. Ioannou, S. T. Liu, H. L. Hughes, „Worst case stress conditions for hot carrier induced degradation of p-channel SOI MOSFETs“, Solid-State Electronics 50 (6), S. 929-934, (2006)

[ITO91] M. Itoh, M. Hori, S. Nadahara, „The origin of stress in sputter-deposited tungsten films for x-ray masks“, Journal of Vacuum Science and Technology B, 9 (1), S. 149-153, (1991)

[JED1] JEDEC/FSA Joint Publication, „Foundry Process Qualification Guidelines“, JP-001, (2002)

[JED122] JEDEC Publication, „Failure Mechanisms and Models for Semiconductor Devices“, JEP122E, (2009)

[JED139] JEDEC Publication, „Constant Temperature Aging to Characterize Aluminum Interconnect Metallization for Stress-Induced Voiding“, JEP139, (2000)

[JED22] JEDEC Standard, „Electrically Erasable Programmable ROM (EEPROM) Program/Erase Endurance and Data Retention Test“, JESD22-A117, (2000)

[JED22b] JEDEC Standard, „High Temperature Storage Life“, JESD22-A103C, (2004)

[JED28] JEDEC Standard, „N-Channel MOSFET Hot Carrier Data Analysis“, JESD28-1, (2001)

[JED28A] JEDEC Standard, „Procedure for Measuring N-Channel MOSFET Hot-Carrier-Induced Degradation Under DC Stress“, JESD28-A, (2001)

[JED35] JEDEC Standard, „Procedure for the Wafer-Level Testing of Thin Dielectrics“, JESD35-A, (2001)

[JED60] JEDEC Standard, „A Procedure for Measuring P-Channel MOSFET Hot-Carrier-Induced Degradation Under DC Stress“, JESD60A, (2004)

[JED61] JEDEC Standard, „Isothermal Electromigration Test Procedure“, JESD61A.01, (2007)

[JED63] JEDEC Standard, „Standard Method for Calculating the Electromigration Model Parameters for Current Density and Temperature“, JESD63, (1998)

[JED87] JEDEC Standard, „Standard Test Structures for Reliability Assessment of AlCu Metallizations with Barrier Materials“, JESD87, (2001)

[JED90] JEDEC Standard, „A Procedure for Measuring P-Channel MOSFET Negative Bias Temperature Instabilities“, JESD90, (2004)

[JED92] JEDEC Standard, „Procedure for Characterizing Time-Dependent Dielectric Breakdown of Ultra-Thin Gate Dielectrics“, JESD92, (2003)

[JEP77] K. O. Jeppson, C. M. Svensson, „Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices“, Journal of Applied Physics, 48 (5), S. 2004-2014, (1977)

[JOH00] C. Johnston, A. Crossley, R. Sharp, „The possibilities for high temperature electronics in combustion monitoring“, Advanced Sensors and Instrumentation Systems for Combustion Processes, IEE Seminar, 2000/080, S. 9/1-9/3, (2000)

[JOH04] R. W. Johnson, J. L. Evans, P. Jacobsen, J. R. Thompson, M. Christopher, „The changing automotive environment: High-temperature electronics“, IEEE Transactions on Electronics Packaging Manufacturing, 27 (3), S. 164-176, (2004)

[KAT08] A. A. Katsetos, „Negative bias temperature instability (NBTI) recovery with bake“, Microelectronics Reliability, 48 (10), S. 1655-1659, (2008)

[KIM03] J. H. Kim, J. B. Choi, B. J. Shin, K. H. Park, „An empirical model for charge leakage through oxide-nitride-oxide interpoly dielectric in stacked-gate flash memory devices“, Semiconductor Science and Technology, 18, S. 158-162, (2003)

[KIM08] J.-J. Kim, R. Rao, S. Mukhopadhyay, C.-T. Chuang, „Ring oscillator circuit structures for measurement of isolated NBTI/PBTI effects“, IEEE conference on integrated circuit design and technology and tutorial, S. 163-166, (2008)

[KIM11] J.-J. Kim, R. Rao, J. Schaub, A. Ghosh, A. Bansal, K. Zhao, B. P. Linder, J. Stathis, „PBTI/NBTI monitoring ring oscillator circuits with on-chip Vt characterization and high frequency AC stress capability“, Symposium on VLSI Circuits Digest of Technical Papers, S. 224-225, (2011)

[KIR12] R. K. Kirschman, http://www.extremetemperatureelectronics.com, (Stand: 19.10.2012)

[KOL86] A. Kolodny, S. T. K. Nieh, B. Eitan, J. Shappir, „Analysis and modeling of floating-gate EEPROM cells“, IEEE Transactions on Electron Devices, 33 (6), S. 835-844, (1986)

[LEB93] Y. Leblebici, S.-M. Kang, „Hot-Carrier reliability of MOS VLSI circuits“, Kluwer Academic Publishers, (1993)

[LEE88] J. C. Lee, I. C. Chen, C. Hu, „Statistical modeling of silicon dioxide reliability“, Proceedings of the International Reliability Physics Symposium, S. 131-138, (1988)

[LIE02] J. Lienig, G. Jerke, „Elektromigration / Eine neue Herausforderung beim Entwurf elektronischer Baugruppen / Teil 1“, F & M Elektronik, Jahrgang 110, S. 36-39, (2002)

[LIE06] J. Lienig, „Introduction to electromigration-aware physical design“, International Symposium on Physical Design ISPD, (2006)

[MAH09] S. Mahapatra, V. D. Maheta, A. E. Islam, M. A. Alam, „Isolation of NBTI stress generated interface trap and hole-trapping components in PNO p-MOSFETs“, IEEE Transactions on Electron Devices, 56 (2), S. 236-242, (2009)

[McP01] J. W. McPherson, „Physics and chemistry of intrinsic time-dependent dielectric breakdown in SiO_2 dielectrics“, International Journal of High Speed Electronics and Systems, 11 (3), S. 751-787, (2001)

[McP10] J. W. McPherson, „Reliability Physics and Engineering“, Springer, Kapitel 11.2, S. 148, (2010)

[McP85] J. W. McPherson, D. A. Baglee, „Acceleration factors for thin gate oxide stressing“, Proceedings of the Reliability Physics Symposium, S. 1-5, (1985)

[MIC01] N. L. Michael, C.-U. Kim, „Electromigration in Cu thin films with Sn and Al cross strips“, Journal of Applied Physics, 90 (9), S. 4370-4376, (2001)

[MIE83] N. R. Mielke, „New EPROM data-loss mechanisms“, Proceedings of the Reliability Physics Symposium, S. 106-113, (1983)

[MOA89] R. Moazzami, J. C. Lee, C. Hu, „Temperature acceleration of Time-Dependent Dielectric Breakdown“, IEEE Transactions on Electron Devices, 36 (11), S. 2462-2465, (1989)

[MOO07] R. Moonen, P. Vanmeerbeek, G. Lekens, W. De Ceuninck, P. Moens, J. Boutsen, „Lifetime modeling of intrinsic gate oxide breakdown at high temperature“, Microelectronics Reliability, 47 (9-11), S. 1389-1393, (2007)

[MOO65] G. E. Moore, „Cramming more components onto integrated circuits“, Electronics, 38 (8), S.114-117 , (1965) oder: G. E. Moore, IEEE Solid-State Circuits Newsletter, 11 (5), S. 33-35, (2006), reprinted from Electronics, 38 (8), S.114-117 , (1965)

[MUL86] R. S. Muller, T. I. Kamins, „Device Electronics for Integrated Circuits“, 2nd Edition, John Wiley & Sons Inc., (1986)

[NOR11] R. Normann, „Synopsis: Report on high temperature tools technologyn from the international partnership for geothermal technology“, HiTEN 2011, (2011)

[OBO65] D. O'Boyle, „Observations on electromigration and the soret effect in tungsten“, Journal of Applied Physics, 36 (9), S. 2849-2853, (1965)

[OGU80] S. Ogura, P. J. Tsang, W. W. Walker, D. L. Critchlow, J. F. Shepard, „Design and characteristics of the lightly doped drain-source (LDD) insulated gate field-effect transistor“, IEEE Transactions on Electron Devices, 27 (8), S. 1359-1367 , (1980)

[OKA93] H. Okabayashi, „Stress-induced void formation in metallization for integrated circuits“, Materials Science and Engineering R, 11 (5), S. 191-241, (1993)

[OSH12] Oshino Lamps, http://www.oshino-lamps.de/prde/tec10.html, (Stand: 15.08.2012)

[OTA12] K. Ota, M. Saitoh, Y. Nakabayashi, T. Ishihara, K. Uchida, T. Numata, „Threshold voltage shift and drain current degradation by negative bias temperature instability in Si (110) p-channel metal-oxide-semiconductor field-effect transistor“, Applied Physics Letters, 100 (21), S. 212109-1 - 212109-3, (2012)

[OVI08] I. A. Ovid'ko, A. G. Sheinerman, E. C. Aifantis, „Stress-driven migration of grain boundaries and fracture processes in nanocrystalline ceramics and metals“, Acta Materialia, 56 (12), S. 2718-2727, (2008)

[PAN91] C.-S. Pan, K. Wu, G. Sery, „Physical origin of long-term charge loss in floating-gate EPROM with an interpoly oxide-nitride-oxide stacked dielectric“, IEEE Electron Device Letters, 12 (2), S. 51-53, (1991)

[PAN95] G. Pananakakis, G. Ghibaudo, R. Kies, C. Papadas, „Temperature dependence of the Fowler-Nordheim current in metal-oxide-degenerate semiconductor structures“, Journal of Applied Physics, 78 (4), S. 2635- 2641, (1995)

[PAP95] C. Papadas, G. Pananakakis, G. Ghibaudo, C. Riva, F. Pio, P. Ghezzi, „Modeling of the intrinsic retention characteristics of FLOTOX EEPROM cells under elevated temperature conditions“, IEEE Transactions on Electron Devices, 42 (4), S. 678-682, (1995)

[PAR03] B. Parmentier, O. Vermesan, L. Beneteau, „Design of high temperature electronics for well logging applications“, HiTEN 2003, (2003)

[PEA68] J. C. Peacock, A. D. Wilson, „Electrotransport of tungsten and life of a filament“, Journal of Applied Physics, 39 (13), S. 6037-6041, (1968)

[PEN94] J. ZZ. Peng, S. Haddad, H. Fang, C. Chang, S. Longcor, B. Ho, Y. Sun, D. Liu, Y. Tang, J. Hsu, S. Luan, J. Lien, „Flash EPROM endurance simulation using physics-based models“, IEEE Electron Device Meeting IEDM, S. 295-298, (1994)

[PER78] G. Perlegos, P. J. Salsbury, „Electrically programmable MOS read-only memory with isolated decoders“, United States Patent, Nr. 4,094,012, (1978)

[REG02] S. Reggiani, M. Valdinoci, L. Colalongo, M. Rudan, G. Baccarani, A. D. Stricker, F. Illien, N. Felber, W. Fichtner, L. Zullino, „Electron and hole mobility in silicon at large operating temperatures - Part I: Bulk mobility“, IEEE Transactions on Electron Devices, 49 (3), S. 490-499, (2002)

[REI06] H. Reisinger, O. Blank, W. Heinrigs, A. Mühlhoff, W. Gustin, C. Schlünder, „Analysis of NBTI degradation- and recovery-behavior based on ultra fast V_T-measurements“, Proceedings of the IEEE International Reliability Physics Symposium, S. 448-453, (2006)

[ROT12] K. Rott, H. Reisinger, S. Aresu, C. Schlünder, K. Kölpin, W. Gustin, T. Grasser, „New insights on the PBTI phenomena in SiON pMOSFETs“, Microelectronics Reliability, 52 (9-10), S. 1891-1894, (2012)

[ROY03] K. Roy, S. Mukhopadhyay, H. Mahmoodi-Meimand, „Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits“, Proceedings of the IEEE, 91 (2), S. 305-327, (2003)

[RYD00] M. Rydberg, U. Smith, „Long-term stability and electrical properties of compensation doped poly-Si IC-resistors“, IEEE Transactions on Electron Devices, 47 (2), S. 417-426, (2000)

[RYD01] M. Rydberg, U. Smith, „Long-term stability and electrical properties of fluorine doped polysilicon IC-resistors“, Materials Science in Semiconductor Processing, 4 (4), S. 373-382, (2001)

[SAK94] E. Sakagami, N. Arai, H. Tsunoda, H. Egawa, Y. Yamaguchi, E. Kamiya, M. Takebuchi, K. Yamada, K. Yoshikawa, S. Mori, „The impact of intermetal dielectric layer and high temperature bake test on the reliability of nonvolatile memory devices“, Proceedings of the IEEE International Reliability Physics Symposium, S. 359-367, (1994)

[SCH94] K. F. Schuegraf, C. Hu, „Metal-oxide-semiconductor field-effect-transistor substrate current during Fowler-Nordheim tunneling stress and silicon dioxide reliability“, Journal of Applied Physics, 76 (6), S. 3695-3700, (1994)

[SCH94b] K. F. Schuegraf, C. Hu, „Reliability of thin SiO_2“, Semiconductor Science and Technology, 9, S. 989-1004, (1994)

[SCH98] D. K. Schroder, „Semiconductor Material and Device Characterization“, 2nd Edition, John Wiley & Sons Inc., (1998)

[SHE06] L. Sheng, E. De Backer, D. Wojciechowski, J. De Greve, K. Dhondt, S. Boonen, D. Malschaert, E. Snyder, „Surface Roughness Enhanced Current in Defectively Stressing Poly-Oxide-Poly Capacitors“, Final report of the IEEE International Integrated Reliability Workshop, S. 205-208, (2006)

[SHI80] R. E. Shiner, J. M. Caywood, B. L. Euzent, „Data retention in EPROMs“, Proceedings of the Reliability Physics Symposium, S. 238-243, (1980)

[SHO84] F. Shoucair, W. Hwang, P. Jain, „Electrical characteristics of large scale integration (LSI) MOSFETs at very high temperatures“, Microelectronics Reliability, 24 (3), S. 465-485, (1984)

[SOM10] S. P. Sommer, Dissertation zum Thema „Plasma Charging Damage bei Bauteilen höchster Zuverlässigkeitsanforderungen“, Universität Duisburg-Essen, (2010)

[STA01] J. H. Stathis, „Physical and predictive models of ultra thin oxide reliability in CMOS devices and circuits“, Proceedings of the IEEE International Reliability Physics Symposium, S. 132-149, (2001)

[STA06] J. H. Stathis, S. Zafar, „The negative bias temperature instability in MOS devices: A review“, Microelectronics Reliability, 46 (2-4), S. 270-286, (2006)

[SUE94] J. S. Suehle, P. Chaparala, C. Messick, W. M. Miller, K. C. Boyko, „Field and temperature acceleration of Time-Dependent Dielectric Breakdown in intrinsic thin SiO_2“, Proceedings of the IEEE International Reliability Physics Symposium, S. 120-125, (1994)

[SUE97] J. S. Suehle, P. Chaparala, „Low electric field breakdown of thin SiO_2 films under static and dynamic stress“, IEEE Transactions on Electron Devices, 44 (5), S. 801-808, (1997)

[SUL92] T. D. Sullivan, „Stress-induced voiding in microelectronic metallization: Void growth models and refinements“, Annual Review of Materials Science, 26, S. 333-364, (1992)

[SUN11] H. Sun, M. Wei, „Stress migration induced formation of voids / hillocks in tungsten films“, Advanced Materials Research, 311-313, S. 1831-1834, (2011)

[SUN93] J. Sune, M. Lanzoni, P. Olivo, „Temperature dependence of Fowler-Nordheim injection from accumulated n-type silicon into silicon dioxide“, IEEE Transactions on Electron Devices, 40 (5), S. 1017-1019, (1993)

[SUZ86] E. Suzuki, D. K. Schroder, Y. Hayashi, „Carrier conduction in ultrathin nitrided oxide films“, Journal of Applied Physics, 60 (10), S. 3616-3621, (1986)

[SZE81] S. M. Sze, „Physics of Semiconductor Devices“, 2nd Edition, John Wiley & Sons Inc., (1981)

[TEZ90] A. Tezaki, T. Mineta, H. Egawa, T. Noguchi, „Measurement of three dimensional stress and modeling of stress induced migration failure in aluminum interconnects“, Proceedings of the International Reliability Physics Symposium, S. 221-229, (1990)

[TSE95] H.-H. Tseng, P. J. Tobin, S. Hong, „Polarity dependence of gate oxide quality with SOI substrates“, Proceedings of the IEEE International SOI Conference, S. 56-57, (1995)

[TU03] K. N. Tu, „Recent advances on electromigration in very-large-scale-integration of interconnects“, Journal of Applied Physics, 94 (9), S. 5451-5473, (2003)

[TUR91] R. Turkman, „Electrical Characterization of polysilicon surface roughness in double polysilicon EPROMs“, Proceedings of the IEEE Microelectronics Symposium, S. 79-83, (1991)

[VAN08] D. Vanhoenacker-Janvier, M. El Kaamouchi, M. Si Moussa, „Silicon-on-insulator for high-temperature applications“, IET Circuits Devices Systems, 2 (1), S. 151-157, (2008)

[VER88] G. Verma, N. Mielke, „Reliability performance of ETOX based flash memories“, Proceedings of the International Reliability Physics Symposium, S. 158-166, (1988)

[VIN96] E. Vincent, N. Revil, C. Papadas, G. Ghibaudo, „Electric field dependence of TDDB activation energy in ultrathin oxides“, Microelectronics Reliability, 36 (11/12), S. 1643-1646, (1996)

[VIR02] Virginia Semiconductor, „The general properties of Si, Ge, SiGe, SiO_2, Si_3N_4“, (2002);
http://www.virginiasemi.com/pdf/generalpropertiessi62002.pdf

[VOL11] C. Volkmer, Bachelorarbeit zum Thema „Entwicklung eines Messprogramms für Hot Carrier Messungen an Feld-Effekt-Transistoren (MOSFET) auf Silicon-On-Insulator (SOI)“, Fachhochschule Düsseldorf und Fraunhofer IMS, (2011)

[WAL92] D. T. Walton, H. J. Frost, C. V. Thompson, „Development of near-bamboo and bamboo microstructures in thin-film strips“, Applied Physics Letters, 61 (1), S. 40-42, (1992)

[WAL96] A. Walter, Dissertation zum Thema „Prozessintegration von EEPROM-Modulen für ASIC-CMOS Technologien“, Universität Duisburg, (1996)

[WAN93] H.-J. Wann, J. King, J. Chen, P. K. Ko, C. Hu, „Hot-carrier currents of SOI MOSFETs“, Proceedings of the IEEE International SOI Conference, S. 118-119, (1993)

[WEN09] Z. Wenbin, C. Haifeng, X. Zhiqiang, L. Leilei, Y. Zongguang, „W-plug via electromigration in CMOS process“, Journal of Semiconductors, 30 (5), S. 056001-1 - 056001-4, (2009)

[WER96] R. Werner, Dissertation zum Thema „Untersuchung einer CMOS-Technologie auf SIMOX-Substraten für die Anwendungen in der Hochtemperaturelektronik“, Universität Duisburg, (1996)

[WES85] N. H. E. Weste, K. Eshraghian, „Principles of CMOS VLSI Design“, Addison-Wesley Publishing Company, (1985)

[WIT00] C. Witt, Dissertation zum Thema „Electromigration in Bamboo Aluminum Interconnects“, Universität Stuttgart, (2000)

[WU01] E. Y. Wu, J. M. McKenna, W. Lai, E. Nowak, A. Vayshenker, „The effect of change of voltage acceleration on temperature activation of oxide breakdown for ultrathin oxides“, IEEE Electron Device Letters, 22 (12), S. 603-605, (2001)

[WU90] K. Wu, C.-S. Pan, J. J. Shaw, P. Freiberger, G. Sery, „A model for EPROM intrinsic charge loss through oxide-nitride-oxide (ONO) interpoly dielectric“, Proceedings of the International Reliability Physics Symposium, S. 145-149, (1990)

[YAN04] B. L. Yang, P. T. Lai, H. Wong, „Conduction mechanisms in MOS gate dielectric films“, Microelectronics Reliability, 44 (5), S. 709-718, (2004)

[YAS99] A. Yassine, H. E. Nariman, K. Olasupo, „Field and temperature dependence of TDDB of ultrathin gate oxide“, IEEE Electron Device Letters, 20 (8), S. 390-392, (1999)

[YEO01] Y.-C. Yeo, Q. Lu, C. Hu, „MOSFET gate oxide reliability: Anode hole injection model and its applications“, International Journal of High Speed Electronics and Systems, 11 (3), S. 849-886, (2001)

Abkürzungsverzeichnis

AEC	Automotive Electronics Council
Akk.	Akkumulation
BD	Breakdown
BG	Bandgap-Referenz-Schaltung
BOX	Buried Oxide
BPSG	Bor-Phosphor-Silikat-Glas
CMOS	Complementary Metal Oxide Semiconductor
CVD	Chemical Vapor Deposition
DIN	Deutsches Institut für Normung
DRAM	Dynamic Random-Access Memory
e^-	Elektron
EEPROM	Electrically Erasable Programmable Read Only Memory
EPROM	Erasable Programmable Read Only Memory
h^+	Loch
H035	Hochtemperatur-SOI-CMOS-Prozess des Fraunhofer IMS mit l = 0,35 µm minimaler Kanallänge
H10	Hochtemperatur-SOI-CMOS-Prozess des Fraunhofer IMS mit l = 1 µm minimaler Kanallänge
HCI	Hot Carrier Injection oder Hot Carrier Integrity
ICS	Interactive Characterization Software
IMS	Fraunhofer-Institut für Mikroelektronische Schaltungen und Systeme
Inv.	Inversion
ISO	International Organization for Standardization

IU-Kennlinien	Strom-Spannungs-Kennlinien
JEDEC	Joint Electron Devices Engineering Council
JESD	JEDEC-Standard
JP, JEP	JEDEC-Publikation
LabVIEW	Laboratory Virtual Instrumentation Engineering Workbench
LB	Leitungsband
LDD	Lightly Doped Drain
MIL	Abkürzung für US-amerikanische Militär-Standards
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
NBTI	Negative Bias Temperature Instability
NMOS	n-channel Metal Oxide Semiconductor
Nplus	Phosphordotierung (ca. 10^{19} 1/cm³)
n-Si	n-dotiertes Silizium
PMOS	p-channel Metal Oxide Semiconductor
Poly, Poly-Si	Polysilizium
Pplus	Bordotierung (ca. 10^{19} 1/cm³)
ppm	one part per million
PROM	Programmable Read Only Memory
p-Si	p-dotiertes Silizium
QBD	Charge to Breakdown
ROM	Read Only Memory
RT	Raumtemperatur
RTA	Rapid Thermal Annealing
SILC	Stress Induced Leakage Current(s)
SOI	Silicon-on-Insulator
SRAM	Static Random-Access Memory

TDDB	Time-Dependent Dielectric Breakdown
TVS	Triangular Voltage Sweep
UV	Ultraviolett
VB	Valenzband
WLR	Wafer Level Reliability

Formelzeichen

C_{depl}	Depletion-Kapazität
C_G	Kapazität zwischen dem Control-Gate und dem Floating-Gate
C_L	Ausgangskapazität eines Inverters
C_{ox}	Kapazität des Gateoxids
D	Diffusivität von Leerstellen im Metall
D_0	Konstante bzgl. der Diffusivität von Leerstellen im Metall
d_{gox}	Gateoxiddicke
D_n	Diffusionskoeffizient für Elektronen
D_p	Diffusionskoeffizient für Löcher
d_{si}	Dicke des Siliziumfilms
d_{tox}	Tunneloxiddicke
e	Elementarladung
E_A	Aktivierungsenergie
$E_{A(1/E)}$	Aktivierungsenergie im 1/E-Modell
$E_{A(E)}$	Aktivierungsenergie im E-Modell
E_c	Energie der niedrigsten Energiestufe des Leitungsbands
E_e	Energie, wenn ein zusätzliches Elektron beschleunigt wird (Hot Carrier)
E_g	Bandlücke von Silizium
E_{ox}	Elektrisches Feld über dem Gateoxid
F	kumulierte Ausfallwahrscheinlichkeit
f	Frequenz
$f(0)$	Frequenz zum Zeitpunkt $t = 0$

G	Feldbeschleunigungsfaktor im 1/E-Modell
g_{m}	Steilheit (Transkonduktanz) der Eingangskennlinie
$g_{\mathrm{m,max}}$	maximale Steilheit (Transkonduktanz) der Eingangskennlinie
h	Plancksches Wirkungsquantum
I	Stromaufnahme von Ringoszillator oder Bandgap-Referenz
$I(0)$	Stromaufnahme von Ringoszillator oder Bandgap-Referenz zum Zeitpunkt $t = 0$
$I_{\mathrm{B,max}}$	maximaler Substratstrom
I_{BD}	Strom beim Oxiddurchbruch
I_{D}	Strom einer Diode in Durchlassrichtung
I_{DS}	Drain-Source-Strom
$I_{\mathrm{DS}}(0)$	Drain-Source-Strom zum Zeitpunkt $t = 0$
$I_{\mathrm{DS,lin}}$	linearer Drain-Source-Strom
$I_{\mathrm{DS,sätt}}$	Sättigungs-Drain-Source-Strom
$I_{\mathrm{G,max}}$	maximaler Gatestrom
I_{leck}	gesamter Leckstrom eines Transistors
$I_{\mathrm{pn,np}}$	Leckstrom der sperrgepolten pn- bzw. np-Übergänge
I_{poly}	am Polysilizium-Gate gemessener Strom
I_{PTAT}	Strom, der linear mit der Temperatur ansteigt (Bandgap-Referenz)
I_{S}	Sättigungssperrstrom einer Diode
$I_{\mathrm{sätt}}$	Sättigungsstrom eines Transistors
I_{sub}	Subtreshold-Leckstrom
ΔI_{DS}	Differenz des Drain-Source-Stromes zwischen $t = t_1$ und $t = t_2$
j	Stromdichte
J_{FN}	Fowler-Nordheim-Stromdichte
k	Boltzmann-Konstante

l	Länge (z.B. Kanallänge eines Transistors)
l_n	Kanallänge eines NMOS-Transistors
l_p	Kanallänge eines PMOS-Transistors
m_{ox}	effektive Elektronenmasse im Oxid
m_{si}	effektive Elektronenmasse im Silizium
$N_{a,d}$	Kanaldotierung von NMOS bzw. PMOS
n_B	Subtreshold-Swing-Koeffizient bzw. Body-Effekt-Koeffizient
n_i	intrinsische Ladungsträgerkonzentration
$P(0)$	Parameter zum Zeitpunkt $t = 0$
$P(t)$	Parameter zum Zeitpunkt t
q	Ladung eines Elektrons bzw. Lochs
Q_{BD}	Ladung bis zum Durchbruch
q_{BD}	Ladung bis zum Durchbruch pro cm^2
Q_{depl}	Ladung im Verarmungsbereich
Q_F	Ladung auf dem Floating-Gate
Q_{ox}	Ladung im Gateoxid
R	Widerstand
$R_\square$	Square-Widerstand
$R(0)$	Widerstand zum Zeitpunkt $t = 0$
$R(t)$	Widerstand zum Zeitpunkt t
T	Temperatur
t	Zeit, je nach Zusammenhang in s oder in h
t_0	Zeitkonstante
$t_{63\%}$	Ausfallzeit für den Weibullwert $W = 0$, d. h. für 63,21 % der Proben
t_{BD}	Zeit bis zum Durchbruch
$t_{d,n}$	Verzögerungszeit des Signals pro Inverter (Ringoszillator, NMOS leitet)

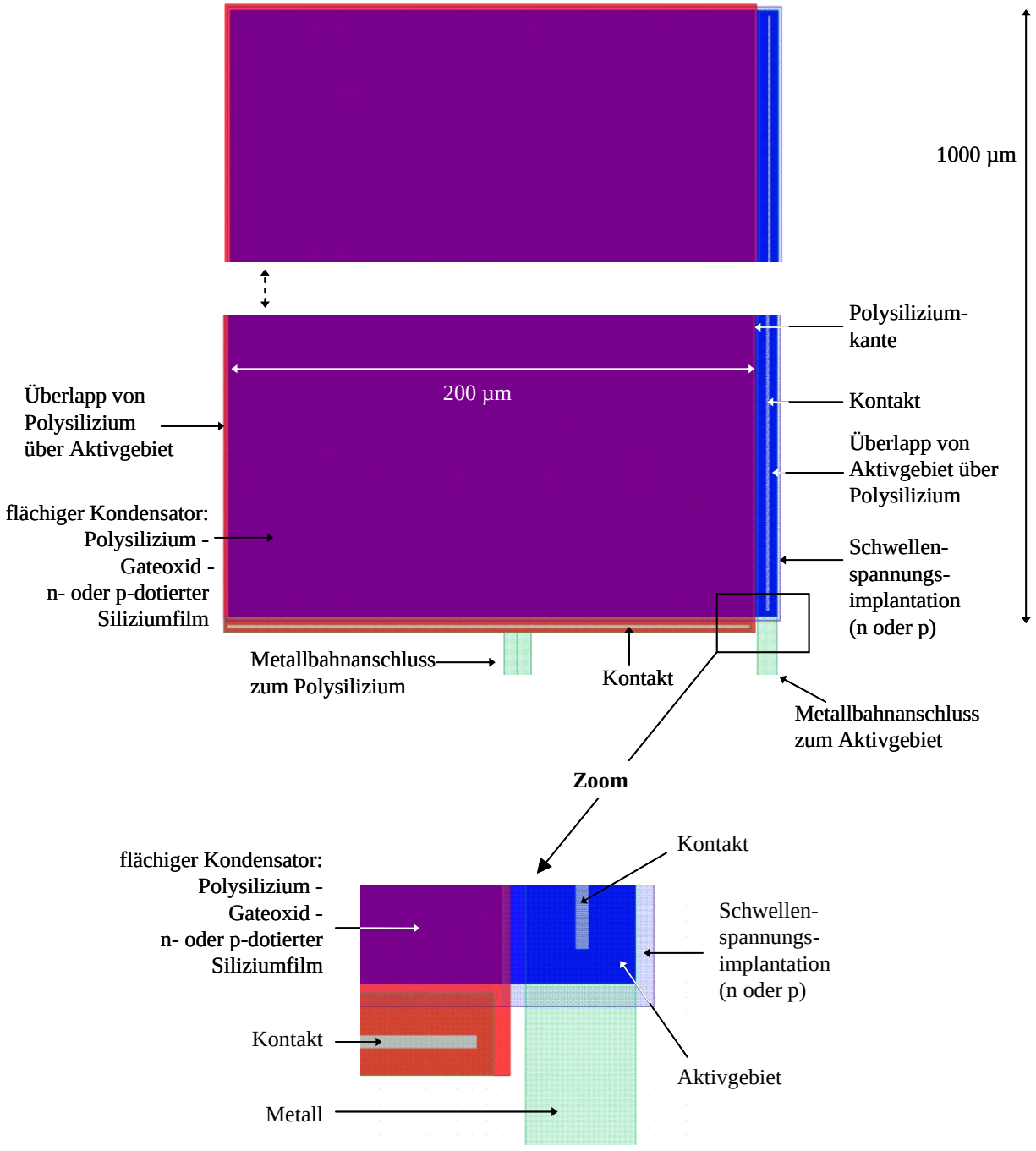
$t_{d,p}$	Verzögerungszeit des Signals pro Inverter (Ringoszillator, PMOS leitet)
T_e	effektive Temperatur
t_{fail}	Zeit bis zum Ausfall (eines Bauelements)
t_{tar}	Zeit bis zum Erreichen einer festgelegten Änderung
ΔT	Temperaturdifferenz (z. B. bei der Stressmigration zwischen der Abscheidetemperatur der Passivierung und der Lagerungstemperatur)
U	angelegte Spannung
U-	negative angelegte Spannung am Polysilizium-Gate
U+	positive angelegte Spannung am Polysilizium-Gate
U_1	Eingangsspannung eines Inverters
U_2	Ausgangsspannung eines Inverters
U_{BD}	Durchbruchspannung des Gateoxids
U_{BG}	Back-Gate-Spannung
U_{BGS}	Back-Gate-Source-Spannung
U_{CG}	Control-Gate-Spannung
U_{Chuck}	Chuckpotential
U_D	Drainspannung
U_{DBG}	Drain-Back-Gate-Spannung
$U_{D,lin}$	linearer Anteil der Diodenspannung
$U_{D,nlin}$	nichtlinearer Anteil der Diodenspannung
U_{dd}	Versorgungsspannung
U_{Diode}	Flussspannung einer Diode
U_{DS}	Drain-Source-Spannung
$U_{DS,max}$	maximal erlaubte Drain-Source-Spannung im Betrieb
$U_{DS,stress}$	Drain-Source-Stressspannung
U_{Dsat}	Sättigungs-Drainspannung

$U_{Eg,0}$	Spannungsanteil einer Diode, der von der Bandlücke bei 0 K bestimmt wird
U_{FB}	Flachbandspannung
U_G	Gatespannung
U_{GS}	Gate-Source-Spannung
$U_{GS,max}$	maximal erlaubte Gate-Source-Spannung im Betrieb
$U_{GS,stress}$	Gate-Source-Stressspannung
U_{ox}	über dem Oxid abfallende Spannung
U_{poly}	am Polysilizium-Gate angelegte Spannung
U_{pp}	Programmierspannung
U_{PTAT}	Spannung, die linear mit der Temperatur ansteigt (Bandgap-Referenz)
U_{ref}	Referenz-Spannung einer Bandgap-Schaltung
U_S	Sourcespannung
U_{SG}	Select-Gate-Spannung
U_{th}	Schwellenspannung
$U_{th}(0)$	Schwellenspannung zum Zeitpunkt $t = 0$
$U_{th,0}$	Schwellenspannung bei Raumtemperatur
$U_{th,ge}$	Schwellenspannung einer gelöschten EEPROM-Zelle
$U_{th,n}$	Schwellenspannung eines NMOS-Transistors
$U_{th,pr}$	Schwellenspannung einer programmierten EEPROM-Zelle
$U_{th,p}$	Schwellenspannung eines PMOS-Transistors
ΔU_D	Differenz der Spannungen zweier Dioden
ΔU_{th}	Programmierfenster (EEPROM-Zellen) oder Änderung der Schwellenspannung (NBTI)
W	kumulierte Anzahl der Ausfälle W (Weibullwert)
w	Weite (z. B. Kanalweite eines Transistors)
w_n	Kanalweite eines NMOS-Transistors

w_p	Kanalweite eines PMOS-Transistors
x_{dmax}	maximale Tiefe der Verarmungsschicht
$Y(t)$	prozentuale Änderung eines Parameters P(t) in Bezug auf P(0)
$Z_{1\,V}$	Zyklenzahl nach 1 V Abnahme des anfänglichen Programmierfensters
$Z_{15\%}$	Zyklenzahl nach 15 % Abnahme des anfänglichen Programmierfensters
β	Steigung einer Geraden im Weibullgraphen
γ	Feldbeschleunigungsfaktor im E-Modell
ε_{si}	Dielektrizitätskonstante von Silizium
μ	Ladungsträgerbeweglichkeit
μ_0	Ladungsträgerbeweglichkeit bei der Bezugstemperatur T_0
μ_n	Ladungsträgerbeweglichkeit von Elektronen
μ_p	Ladungsträgerbeweglichkeit von Löchern
τ_n	Lebensdauer eines Lochs
τ_p	Lebensdauer eines Elektrons
Φ_b	Potentialbarriere vom Polysilizium zum Siliziumdioxid
Φ_F	Fermipotential
Φ_{MS}	Differenz der Austrittsarbeiten zwischen Gate- und Filmmaterial
Φ_S	Oberflächenpotential

Anhang

A) Teststrukturen für die TDDB-Messungen

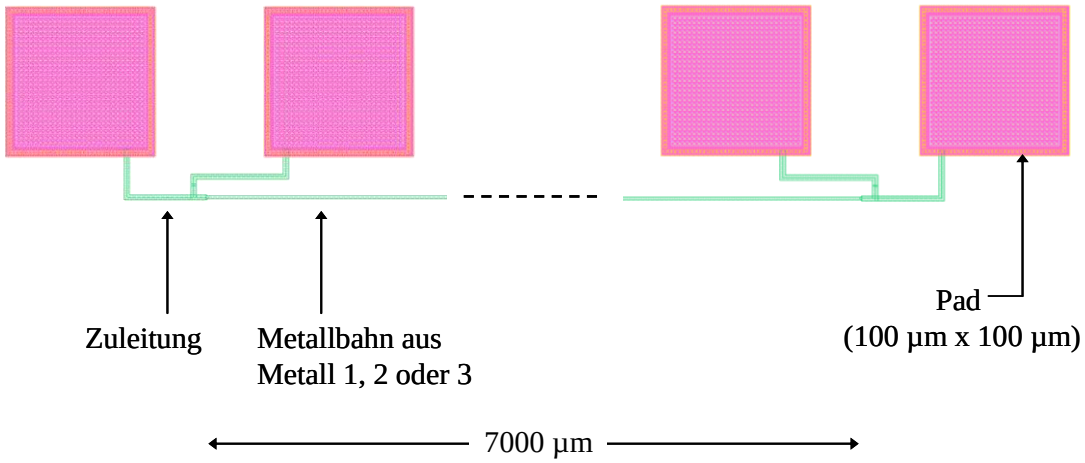


Kondensatoren (A = 200 µm x 1000 µm); flächig Poly über n- oder p-dotiertem Aktivgebiet

B) Teststrukturen für die Elektromigrations- und Stressmigrations-messungen

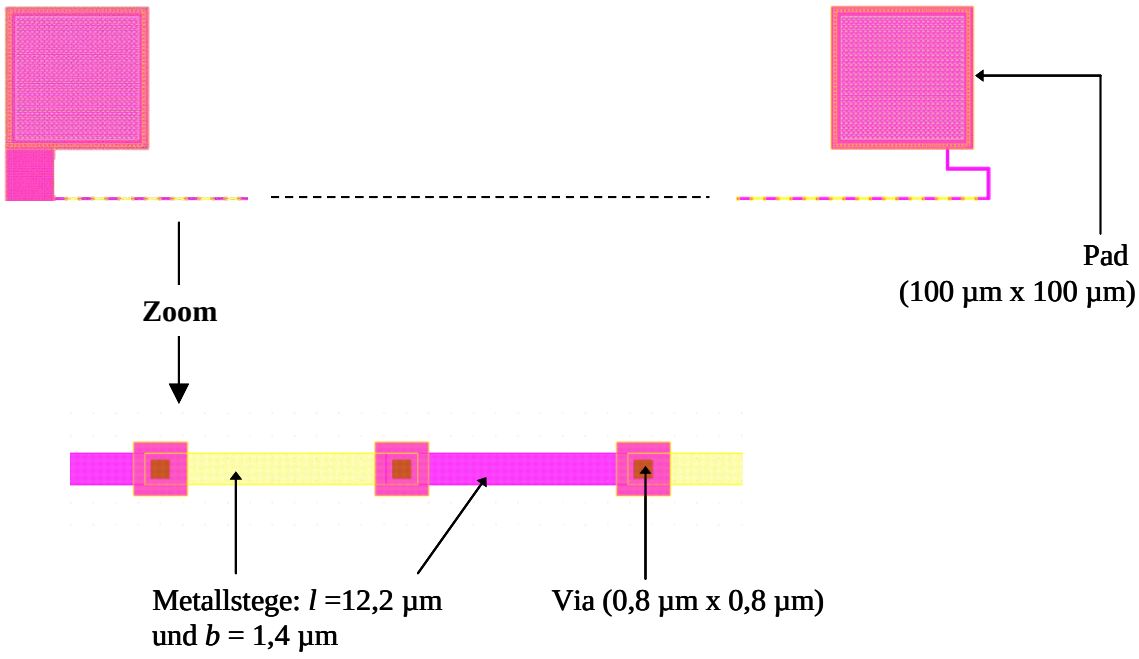
1)

Metallbahn (l = 7000 µm, b = 2 µm)



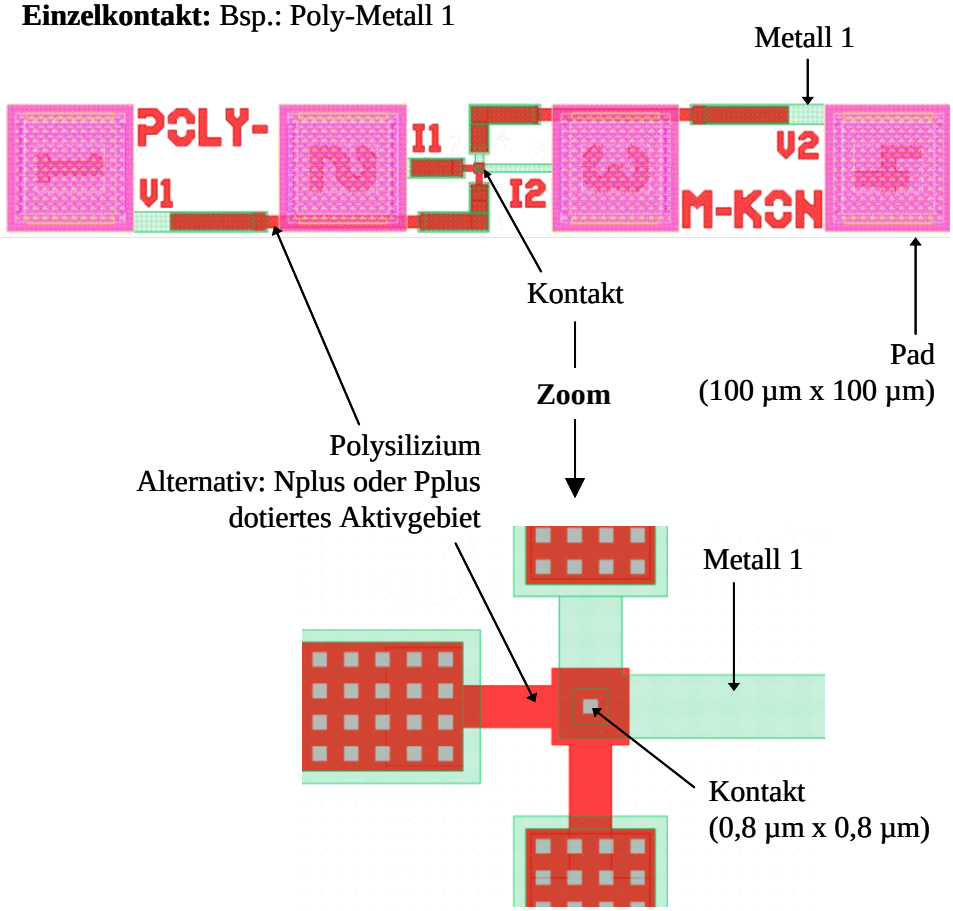
2)

Viakette 100 „Via 2“ zwischen Metall 2 - und Metall 3 - Stegen
oder
100 „Via 1+2“ (übereinander) zwischen Metall 1 - und Metall 3 - Stegen



3)

Einzelkontakt: Bsp.: Poly-Metall 1



4)

Kontaktkette: Bsp.: Poly-Metall 1